Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék

Spontán magyar nyelvű beszéd gépi felismerése nyelvspecifikus szabályok nélkül

Doktori értekezés

Mihajlik Péter
okl. villamosmérnök

PhD fokozat elnyeréséért
BME-VIK Villamosmérnöki Doktori Iskola

Budapest, 2010.

Tartalomjegyzék

Abstract

of the PhD Thesis of Péter Mihajlik,
"Recognition of Spontaneous Hungarian Speech without Language Specific Rules"

Hungarian language large vocabulary continuous speech recognition results are introduced in several successive chapters. The main objective of the dissertation is to investigate whether language specific rules and expert knowledge are indispensable for Hungarian speech recognition, particularly for spontaneous speech. The first chapter briefly summarizes the fundamentals of automated speech recognition. In the second chapter, the weighted finite state transducer framework and its application in speech recognition is surveyed. The third chapter introduces the research objectives of the PhD thesis. The main parts of the study are the next four chapters where the original contributions are detailed.

In the fourth chapter the modeling of the interaction of adjacent speech sounds is discussed. Two explicit phonetic coarticulation modeling techniques are compared with each other and with the implicit – context-independent – baseline. The more data-driven explicit method provides significantly better results than the other one. Furthermore, both explicit modeling approaches outperform drastically the context-independent technique applied widely for Hungarian.

The fifth chapter deals with Hungarian language phonological coarticulation modeling from the perspective of the automatic speech recognition. The results show that the explicit (unweighted) modeling of typical phonological phenomena is not essential for Hungarian speech recognition.

In the sixth chapter a subword lexical modeling technique is proposed for large vocabulary continuous recognition of spontaneous Hungarian speech. In our approach morpheme-like units – called as morphs – form the basis of language and pronunciation models. Statistically derived morphs provided significantly higher speech recognition accuracies than words in the spontaneous language experiments. Moreover, statistical morphs achieved competitive result as compared to that of the grammatical morphs applied in the experiments.

In the seventh chapter the classical phoneme-based acoustic modeling is compared with grapheme-based ones in Hungarian language spontaneous speech recognition tasks. The classical phoneme-based approach requires several language specific rules and expert linguistic knowledge (e.g., weighted pronunciation alternatives, exception dictionaries, grapheme-to-phoneme rules, linguistic classification of speech sounds) whereas in the grapheme-based technique none of them is needed. The fully data driven grapheme-based speech recognition approach – that did take Hungarian morphology into account through statistical morph language units – achieved competitive results as compared to the classical word-phoneme baseline. Moreover, it outperformed significantly the baseline in terms of some important evaluation metrics.

Investigating large vocabulary continuous speech recognition of spontaneous Hungarian, it can be concluded that a fully data driven technique using no language specific rules or expert knowledge can be not only more effective in terms of development time and cost than the classical approach, but also competitive in terms of speech recognition accuracy.

Kivonat

Mihajlik Péter „Spontán magyar nyelvű beszéd gépi felismerése nyelvspecifikus szabályok nélkül” című PhD értekezéséhez

Az értekezés egymásra épülő fejezeteken és általánosságra törekvő magyar nyelvű gépi beszédfelismerési feladatokon keresztül mutatja be a szerző tudományos munkáját, támasztja alá téziseit. Az első fejezet áttekinti a klasszikus gépi beszédfelismerés alapfogalmait, majd a második fejezet a súlyozott véges állapotú átalakítók beszédfelismerési alkalmazását ismerteti. Ezután az értekezés célkitűzési következnek. A tanulmány törzsében alapvetően négy beszédfelismerési szint vizsgálata történik, melyek az első négy téziscsoportnak feleltethetőek meg.

A negyedik fejezet a beszédhangok egymásra hatásának modellezésével foglalkozik (I. téziscsoport). A fonetikai koartikuláció explicit modellezésére két módszert tárgyal és vet össze az implicit, környezetfüggetlen modellezéssel. A nagyobb mértékben adatvezérelt explicit módszer szignifikánsan magasabb felismerési pontosságot biztosít mint a másik, ugyanakkor mindkét explicit megközelítés drasztikusan csökkenti a beszédfelismerési hibát az implicit – a magyar nyelvre máig gyakorta alkalmazott – környezetfüggetlen megközelítéshez képest.

Az ötödik fejezet a magyar nyelvű fonológiai koartikuláció, azaz a hasonulási szabályok szóhatárokon is átívelő modellezésének kérdésével foglalkozik a gépi beszédfelismerés szempontjából (II. téziscsoport). A kísérletek tanúsága szerint a tipikus hasonulási jelenségek komolyabb elméleti hátteret és eszköztárat igénylő (súlyozatlan) explicit modellezése nem nélkülözhetelen a magyar nyelvű gépi beszédfelismerésben.

A hatodik fejezet javaslatot tesz a magyar nyelvű spontán nagyszótáras beszéd felismerésénél a morfológiai változatosság kezelésére (III. téziscsoport). A javasolt megoldásban szavak helyett morfémaszerű egységek képzik mind a nyelvi, mind az akusztikai-kiejtési modellek alapjait. A spontán magyar nyelvű beszédfelismerési kísérletekben a statisztikai alapokon származtatott „morf” lexikai egységekkel szignifikánsan nagyobb felismerési pontosságot sikerült elérni, mint a hagyományos szó egységekkel. Továbbá, a nyelvi elemzővel előállított morféma egységekkel elértekhez viszonyítva is versenyképes eredmények születtek.

A hetedik fejezetben a klasszikus fonéma alapú kiejtési és akusztikus modellezést vetjük össze az ún. graféma alapú modellezéssel spontán magyar nyelvű beszédfelismerési feladaton (IV. téziscsoport). A klasszikus megoldásnál számos nyelvspecifikus szabályra van szükség (valószínűségi súlyozású alternatív kiejtések, idegen és hagyományos írású szavak kivételes kiejtései, graféma-fonéma átalakítási szabályok, a beszédhangok fonetikai osztályozása), míg graféma alapon ezeket mind nélkülözni lehet. A teljesen adatvezérelt, de a magyar nyelv morfológiai jellegét – statisztikai morf lexikai modellek révén – figyelembe vevő beszédfelismerési megközelítés nemcsak versenyképesnek bizonyult a klasszikus szó-fonéma alapú technikával szemben, de fontos metrikák szerint szignifikánsan jobban is teljesített.

Az értekezésben a gépi beszédfelismerés korábban nyelvspecifikusnak tekintett vetületeit vizsgálva arra a végkövetkeztetésre juthatunk, hogy az adott beszédfelismerési feladat esetén nyelvspecifikus szakértői szabályok nélkül, pusztán adatvezérelt technikákat alkalmazva nemcsak egyszerűbben és gazdaságosabban, de kompetitív minőségben is elvégezhető a beszéd-szöveg átalakítás (V. téziscsoport).

Köszönetnyilvánítás

Ezúton szeretném megköszönni témavezetőimnek, Tatai Péter címzetes egyetemi docensnek és Dr. Gordos Géza emeritusz professzornak mindazon sokrétű segítségüket mellyel a PhD fokozatért folyó munkám során kísértek. Köszönöm Dr. Szarvas Máténak szakmai és baráti támogatását, mely a gépi beszédfelismerési kutatásaimat megalapozta. Köszönettel tartozom Fegyó Tibornak a több mint egy évtizednyi közös munkáért, az eredmények hasznosíthatóvá tételéért. Hálával tartozom minden szerzőtársamnak és a TMIT tanszéken minden kollégámnak, aki munkámat figyelemmel követte. Itt köszönöm meg Dr. Takács Györgynek (PPKE) és Dr. Tóth Lászlónak (SZTE-MTA), hogy értékes észrevételeikkel segítettek az értekezés jobbá tételében. Végül, de nem utolsósorban, köszönetet mondok Dr. Németh Gézának a beszédlabor vezetéséért, Dr. Vicsi Klára Professzor asszonynak a beszédadatbázisok rendelkezésre bocsátásáért, illetve Prof. Dr. Sallai Gyula tanszékvezető úrnak az ösztönzéséért.

Az új tudományos eredmények elérését a következő, a magyar állam és/vagy az európai unió, valamint az ipar, illetve magánszemélyek által támogatott kutatás-fejlesztési projektek is elősegítették: NKFP-2/034/2004, GVOP-3.1.1-2004-05-0385/3.0, OM-00102-2007, KMOP-1.1.1-07/1-2008-0034, and by TAMOP-4.2.2/08/1/KMR, NSF ITR Grant No IIS-0122466.

1. A gépi beszédfelismerés alapmódszerei

1.1. Bevezetés

A gépi beszédfelismerés kutatása több évtizedes múltra tekint vissza nemzetközi és hazai viszonylatban is. Az első gyakorlatban is használható módszer a DTW (Dynamic Time Warping, dinamikus idővetemítés) volt [Vintsjuk 68],[Myers & Rabiner 81], [Gordos & Takács, 83]. Ez a dinamikus programozáson alapuló egyszerű eljárás elsősorban nyelvtől független kis szótáras, személyfüggő beszédfelismerésre használható. Lényege, hogy tárolt referenciamintákhoz hasonlítja a bejövő beszédjel lényegkiemelt változatát, és a legjobban illeszkedő referenciamintára, mint felismerési eredményre dönt. Jelentősen korlátozza a megközelítés gyakorlati alkalmazhatóságát, hogy a felhasználónak kell betanítania a rendszert a referencia felvételek egyenkénti bemondásával, és így a felismerő beszélőfüggő lett.

A természetes igény, hogy a gépi beszédfelismerő képességei ne korlátozódjanak egy ember hangjára, maga után vonta a statisztikai szemlélet megjelenését, mely a rejtett Markov-modell keretrendszerben vált elterjedté [Baker 75], [Jelinek & Bahl+ 75]. Azóta is a legjobban bevált módja a beszélőfüggetlen felismerésnek, hogy sok ember beszédét (ill. szövegét) rögzítjük, és e sokaság alapján alkotjuk meg például akusztikus (ill. nyelvi) modelljeinket, melyek a beszéd adott szakaszait (szavait, beszédhangjait, vagy beszédhangrészleteit, ill. szósorozatait) jellemzik.

Természetesen a statisztikai modelleken alapuló felismerési folyamat messzemenőkig nem tévedhetetlen, hiszen egy véges tanító-mintahalmaz alapján próbálunk meg következtetéseket levonni egy gyakorlatilag végtelen teszthalmazra vonatkozóan. Mégis – ahogy ez az értekezés is mutatja –, az egyéb, pl. szakértői szabály alapú megközelítések mind inkább háttérbe szorulnak a beszédfelismerési alapmódszerek között. A következőkben röviden áttekintjük a gépi beszédfelismerésnél általánosan alkalmazott modellezési technikákat, megközelítéseket.

1.2. Az általános beszédfelismerési folyamat

Amikor általánosságban (automatikus) beszédfelismerésről beszélünk, ez alatt a beszéd hangnyomás – idő függvényének gépi módszerekkel történő szöveggé átalakítást értjük.[1]

Az általános beszédfelismerési folyamat – hasonlóan a legtöbb összetett felismerési feladathoz (pl. beszélő-, kézírás-, ujjlenyomat-, optikai karakterfelismerés) – az alábbi két fő lépésre bontható.

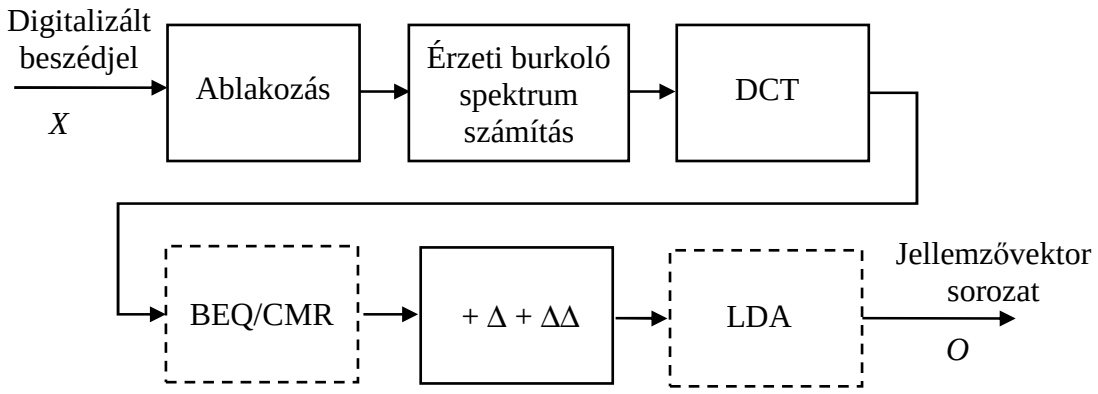
1. Lényegkiemelés
2. Mintaillesztés

1.2.1. Lényegkiemelés

A lényegkiemelés során olyan adatok kiemelésére törekszünk, melyekből jól következtethetünk a felismerni kívánt tartalmi információra. Esetünkben azok a beszédjelből számítható speciális akusztikai beszédjellemzők merülnek fel, melyek a hangképző (beszélő) szervek állásával korrelálnak. A legjobban a rövid idejű spektrális burkoló görbe érzeti

[1] Teljesen általános esetben az audiovizuális gépi beszédfelismerésről is szólnunk kellene, azonban jelen tanulmány keretei között csak annyit jegyezhetünk meg, hogy a vizuális információ megfelelő hozzáadása az audio jelhez jótékonyan befolyásolhatja a felismerési eredményeket (különösen nagy háttérzaj esetén) lásd pl. [Basu & Neti+ 99], [Nefian & Liang+ 02], [Czap 05].

transzformációján alapuló eljárások váltak be (pl. MFCC: Mel Frequency Cepstral Coefficients, PLP: Perceptual Linear Prediction, stb) [Mermelstein 76], [Hermansky 90], ezek kimenete alapesetben a 10-20 együtthatóból álló paramétervektor.



1.1. Ábra. A beszédfelismerési lényegkiemelés tipikus általános blokksémája. (Szaggatott vonallal a nem kötelező elemeket jelöltük.)

A különféle zajok - elsősorban a konvolúciós zajok (pl. csatornatorzítás) - hatását mérséklendő további transzformációs lépések alkalmazhatók. Ezek közül a leggyakrabban a kepsztrális átlagkivonást (CMR), és a vak csatornakiegyenlítést (BEQ) használják [Mauuary 98], melyek közül az első inkább off-line, míg a második on-line módon is számolható.

Ezeken túl fontos megemlíteni, hogy a lényegkiemelésnél nem csupán a statikus (a hangképző szervek pillanatnyi állásával korreláló) paramétereket, hanem azoknak lineáris regressziós becsléssel számított időbeli deriváltjait (sebesség: Δ és gyorsulás: $\Delta\Delta$ paraméterek) is használni szokták. Így a beszéd dinamikája is bizonyos mértékben megjeleníthető a lényegkiemelési kimenetben. Egyes lényegkiemelőkben lineáris diszkriminánsanalízist (LDA) [Haeb-Umbach & Ney 92] is alkalmaznak a lényegkiemelés végső fázisaként.

A lényegkiemelés kimenete az $O = o_1, o_2, \ldots, o_T$ jellemzővektor-sorozat, ahol a vektorok mérete és időzítése állandó[2]. A dimenziószám tipikusan 30-60 közötti, és a vektorok jellemzően 10 ms-onként követik egymást. A továbbiakban a lényegkiemelési eljárások részleteivel nem foglalkozunk, mivel azok a további feldolgozás számára szinte közömbösek.

1.2.2. Mintaillesztés

Beszédfelismerésnél a mintaillesztés feladata a bejövő jellemzővektor-sorozat alapján becslést adni az elhangzott szósorozatra. Kettős feladatot kell megoldani: az osztályozást – melyik beszédrészlet-modell a legvalószínűbb az adott pillanatban, és az időillesztést – melyik időszegmenst rendeljük az egyik, ill. másik modellhez. Ez úgy valósítható meg hatékonyan, hogy a mintát, azaz a bejövő vektorsorozatot (statisztikai úton becsült) valószínűségi modell struktúrákhoz illesztjük, és a legjobb illeszkedésre döntünk (Viterbi-approximáció). A legjobb illesztés keresését szokták önmagában „keresésnek” vagy „dekódolásnak” nevezni, és az ahhoz tartozó szósorozatot pedig a felismerés eredményének.

[2] Léteznek más, nem rögzített jellemzővektor időzítéssel számoló megközelítések is, pl. a szegmens alapú modellezés [Glass 03], de ezek nem terjedtek el széles körben.

Formálisan:

$$\hat{W} = \arg\max_{W} P(W \mid O) \tag{1.1}$$

ahol $W = w_1,...,w_K$, $K \in N$ egy megengedett (modellezett) szósorozatot jelöl, és $O = o_1,...,o_T$ a bejövő beszédjel T elemű lényegkiemelt vektorsorozatát jelöli. $\hat{W} = \hat{w}_1,...,\hat{w}_{\hat{K}}$, $\hat{K} \in N$ pedig a felismert szósorozatot jelenti.

(1.1) a Bayes-szabály segítségével a következőképpen alakítható át:

$$\hat{W} = \arg\max_{W} P(W) \cdot P(O \mid W) \tag{1.2}$$

A (1.2) egyenletet a beszédfelismerés MAP alapegyenletének is szokták nevezni. A formula szemléletesen választja szét az adott szósorozatnak a nyelv által önmagában becsült valószínűségét, $P(W)$-t, az akusztikai megfigyelés valószínűségétől, $P(O \mid W)$-től. Az előbbi valószínűséget a *nyelvi modell* adja, az utóbbit az *akusztikus modell*.

A beszédfelismerésnél a mintaillesztés feladata tehát nem más, mint adott nyelvi és akusztikus modell, valamint bejövő akusztikus jellemzővektor-sorozat esetén a legnagyobb valószínűségű szósorozat megtalálása.

1.3. Felismerési modellhierarchia

Mint láttuk, a beszédfelismerési mintaillesztési feladat két fő szintre bontható: akusztikus és nyelvi modellezési szintre. Az akusztikus modellezés további hierarchiaszintekre osztható, ahol az egyes szintek nyelvspecifikus tudásforrásokkal (kiejtési szótár, elemi akusztikus modellek stb.) reprezentálhatók.

1.3.1. Nyelvi modell

A nyelvi modell, $P(W)$, vagyis adott W szósorozat önmagában vett valószínűségének becslése tipikusan statisztikai közelítő módszerekkel történik.

A legelterjedtebbek az ún. N-gram modellek. Az N-gram modellek közelítő becslést adnak egy tetszőlegesen hosszú szósorozat valószínűségére. A közelítés lényege, hogy az együttes valószínűség kiszámításához használt láncszabályt módosítva alkalmazzuk. A feltételes valószínűségeknél a feltételben a memóriahossz („history”) maximum N-1 lehet (1.3 és 1.4).

$$P(W) = P(w_1,...,w_K) = P(w_1) \cdot P(w_2 \mid w_1) \cdot ... \cdot P(w_K \mid w_{K-1},...,w_1) \tag{1.3}$$

$$P(W) = P(w_1,...,w_K) \approx P(w_1) \cdot P(w_2 \mid w_1) \cdot ... \cdot P(w_K \mid w_{K-1},...,w_{K-(N-1)}) \tag{1.4}$$

A gyakorlatban főként az *N=2* (bigram) és *N=3* (trigram) nyelvi modelleket használják, melyek – főként az előbbi esetén – könnyen integrálhatók a beszédfelismerő rendszerekbe.

1.3.2. Akusztikus modell

Az akusztikus modellt, $P(O|W)$-t, tradicionálisan fonológiai (szó→fonéma) kiejtési, $P(\Phi|W)$, és fonéma szintű akusztikus modellekre, $P(O|\Phi)$, szokták bontani. Ekkor az (1.2) egyenlet az alábbira módosul (Viterbi-approximációt feltételezve):

$$\hat{W} = \arg\max_{W} P(W) \cdot P(\Phi|W) \cdot P(O|\Phi) \tag{1.5}$$

ahol $\Phi = \varphi_1,...,\varphi_P$ a W szósorozathoz tartozó egy lehetséges fonéma- vagy beszédhangsorozat[3]. Látható, hogy a nyelvi modell és az elemi akusztikai modell szintek közé beékelődött a kiejtési modell, ami adott szósorozat esetén egy adott fonémasorozat realizációjának valószínűséget szolgáltatja.

Fonológiai kiejtési modell: A (fonológai) kiejtési modell az egyes szósorozatokhoz valószínűségekkel ellátott fonémasorozatokat rendel. Ez tipikusan egy kiejtési szótár segítségével történik, melyben az egyes ortografikus szóalakokhoz több, valószínűséggel ellátott fonéma sorozat is tartozhat.

$$P(\Phi|W) = \prod_{k=1}^{K} P\left(\Phi_k^i \mid w_k\right) \tag{1.6}$$

ahol Φ_k^i a Φ k-adik szóhoz tartozó i-ik kiejtési variánsának megfelelő rész-fonémasorozata[4]. A mintaillesztés az ML (Maximum Likelihood) értelemben optimális szó-fonéma részsorozat összerendelést, azaz az optimális kiejtési variáns megtalálását is magában foglalja.

A kiejtési szótár előállítása történhet kézi, részben vagy teljesen automatikus módszerekkel.

Fonéma akusztikus modell: A fonéma szintű akusztikus modell feladata adott (megengedett) fonémasorozathoz valószínűséget rendelni az akusztikai megfigyeléssorozat alapján. Formálisan:

$$P(O|\Phi) = \prod_{p=1}^{P} P\left(O_p \mid \varphi_p\right) \tag{1.7}$$

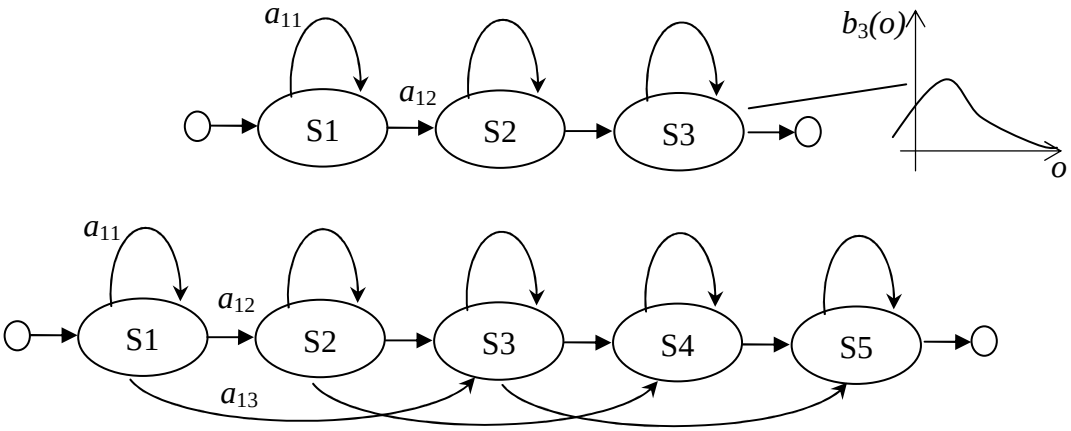
ahol O_p a p-ik fonémához tartozó p-ik jellemzővektor részsorozat. Az (1.5) szerinti maximalizálás tehát itt az optimális fonéma – jellemzővektor részsorozat hozzárendelés megtalálását jelenti.

Feltéve, hogy adott egy (hipotetikus) fonéma modell – jellemzővektor részsorozat hozzárendelés, az akusztikus valószínűségek modellenként számolhatók. A gyakorlatban a fonémák, ill. beszédhangok akusztikai modellezésére a rejtett Markov-modellek (HMM: Hidden Markov-model) használata terjedt el. Általában az ún. „balról-jobbra” struktúrájú, 3 állapotú HMM-eket használnak, de egyes rendszerekben az ún. Bakis struktúra alkalmazása is előfordul [Schramm 06], lásd 1.2. ábra.

---

[3] Az angolban a PLU (Phone Like Unit) elnevezés használatos, mely alatt fonémát és beszédhangot is érthetünk. Jelen értekezésben mind a „fonémák”, mind a „beszédhangok” alatt általában a fizikai beszédhangok absztrakt csoportjait értjük. A fonéma hivatalos definíciójától eltérően a hosszú és rövid mássalhangzókat itt nem különböztetjük meg. A beszédhang-csoportok száma nagyobb lehet, és kialakításuk megközelítéstől függhet.

[4] Adott w_k szóhoz különböző Φ_k^i fonémasorozatok különböző valószínűséggel tartozhatnak, amit táblázatos formával szoktak megadni a nem 0 valószínűségű esetekre.



1.2. ábra. A legegyszerűbb „balról-jobbra” (fent), illetve a Bakis HMM struktúra (lent) beszédhangok akusztikai modellezésére. (A $b_3(o)$ lokális valószínűségsűrűség függvény egy dimenziós jellemzővektorral szemléltetve.)

Adott struktúra és HMM paraméterek esetén a lokális részvalószínűség adott fonéma és jellemzővektor részsorozat esetén:

$$P(O_p \mid \varphi_p) = \max \prod_{t^p=t_0^p}^{T^p} a_{ij}^p b_j^p(o_{t^p}) \tag{1.8}$$

ahol *a* az átmeneti valószínűségeket, *b* az egyes HMM állapotokban értelmezett folyamatos megfigyelési sűrűségfüggvényeket jelenti. (Tehát CD-HMM, azaz Continuous Density Hidden Markov Modell-ekről beszélünk [Levinson & Rabiner+ 83], amit a továbbiakban külön nem jelölünk.)

- *Környezetfüggetlen fonéma akusztikus modellek:*

Ilyenkor a fonéma kiejtési modell a szó kiejtési modellhez hasonlóan környezettő függetlenül szótár formájában is megadható. Fontos, hogy itt a leképezés egyértelmű, azaz egy fonémához pontosan egy elemi akusztikus modell sorozat tartozik. (A továbbiakban a legegyszerűbb balról-jobbra struktúrát feltételezzük.)

- *Környezetfüggő fonéma akusztikus modellek:*

Mivel az egyes beszédhangok artikulációja jelentős mértékben függhet a környező beszédhangoktól, bevett gyakorlat ezt explicit módon modellezni. Általános esetben egy adott kontextusban lévő fonéma környezetfüggő beszédhangmodelljét a hangon felül annak *c* számú jobb és/vagy baloldali szomszédja együttesen határozza meg. A gyakorlatban *c*=1 használata tipikus[5], mindkét szomszédos hang figyelembe vételénél ezeket a beszédhangmodelleket trifónnak, csak az egyik oldali szomszédot tekintve a megfelelő oldali bifónnak, míg a környezetfüggetlen esetet, melyről az előző fejezetben beszéltünk monofónnak nevezzük.

[5] Bár tesznek említést *c*=2,3 környezetfüggésű, azaz „quinphone” és „septphone”-okról, ezek nem terjedtek el széles körben.

1.4. Mintaillesztés rejtett Markov-modellekkel

Megmutatható, hogy a rejtett Markov-modellek bizonyos kiterjesztésével – szószintű kimeneti címkék, ill. nem emittáló állapotok bevezetésével – a fent említett tudásforrások egyetlen (kiterjesztett) rejtett Markov-modellbe integrálhatók (lásd még 2. a fejezetet). A kimeneti címkékkel való kiegészítés azért szükséges, mert nem az elemi (pl. fonéma) rejtett Markov-modellek sorozatára, hanem az ezeknek megfelelő szósorozatra van szükségünk a felismerés eredményeként. A szócímkék beillesztése az egyes elemi modellek közé technikailag könnyen megoldható. Az összetett, kiterjesztett HMM hálózatot a továbbiakban felismerési hálózatnak fogjuk nevezni.

A mintaillesztés HMM alapú felismerés esetén a gyakorlatban nem jelent mást, mint adott bejövő jellemzővektor-sorozat esetén a felismerési hálózatban a ML értelemben legjobb útvonalat megtalálni a kezdőpont és az egyik végpont között. A legjobb útvonal bejárása során érintett szócímkék adják a felismerési eredményét, $\hat{W}$ -t.

1.4.1. Dekódolás

A dekódolás az (1.2) egyenlet optimalizált megoldását, gyakorlatilag a felismerési hálózat legnagyobb valószínűségű útvonalának optimalizált megtalálását („search") jelenti. Optimalizálás alatt a memória és főként a számítási igények leszorítását, kézben tarthatóságát értjük, ahol az előbbiek a felismerési pontosság rovására némi áldozattal is járhatnak.

A dekódolási feladatra a gyakorlatban elterjedten használják a dinamikus programozáson [Bellman 57] alapuló *Viterbi-algortimust*, mely lépésről lépésre képes meghatározni minden diszkrét időpillanatban a HMM felismerési hálózat minden egyes állapotában az odáig tartó legjobb felismerési útvonalat [Ney 84]. Az eljárás jelentős mértékben gyorsítható, amennyiben nem mindegyik részútvonalat számoljuk tovább, csak a legvalószínűbbeket. Ez utóbbi egyes változatait összefoglalóan részleges keresésnek („pruning" [Young 06], „beam search" [Ney & Mergel+ 87]) nevezzük, melynek alkalmazása gyakorlatilag nélkülözhetetlen. A valószínűségi küszöb, ill. egyéb küszöbök megválasztásával lehet beállítani a kompromisszumot a felismerési sebesség és pontosság között.

1.5. Modellparaméterek becslése

Statisztikainak azért nevezik a korszerű beszédfelismerési megközelítéseket, mert az egyes (nyelvi, akusztikai stb.) modellek paramétereinek becslése tanító- vagy mintaadatok (szöveg, hullámforma) alapján, statisztikai úton történik. Minél több és változatosabb a tanítóadat, annál pontosabb és jobb általánosító képességű modelleket készíthetünk. Például, akusztikus modelleknél néhányszor százbeszélős tanító-adatbázisoknál már beszélőfüggetlennek tekinthetjük a betanított modelleket.

1.5.1. Nyelvi modellek

Az N-gram nyelvi modellek paramétereinek becslése tehát statisztikai eljárással történik. A tanító szövegadatbázisban szereplő szó N-esek feltételes relatív gyakorisága ML értelemben becsli a meghatározni kívánt N-gram valószínűségeket.

$$P_{ML}(w_K \mid w_{K-1},...,w_{K-(N-1)}) = \frac{c(w_K, w_{K-1},...,w_{K-(N-1)})}{c(w_{K-1},...,w_{K-(N-1)})} \tag{1.9}$$

ahol a *c(…)* az adott szó N-es, illetve N-1-es előfordulási számát („count") jelöli.

Az N-gram nyelvi modellparaméterek tehát elvileg igen egyszerűen meghatározhatók. A gyakorlatban azonban számos becslési probléma merül fel. Például, akármilyen nagy szövegadatbázist használunk is a tanításhoz, N>1 esetén igen valószínű, hogy lesz olyan szó N-es, ami nem fordul elő a szövegkorpuszban. Ilyen esetekben az (1.9) alapján 0 valószínűséget becslünk az adott N-gram valószínűség számára, ami azt jelenti, hogy az adott N-1 szó után semmiképp nem léphetünk tovább az N-ik szóra. Hasonlóan, problémát okozhat, ha egy szó N-es ugyan előfordul a tanítókorpuszban, de csak kevésszer, és így megfelelő N-gram valószínűség becslése pontatlan lehet.

A fenti adatelégtelenségi („data sparsity”) problémák mindig fellépnek, ha mégsem, akkor nem elég komplex a modellünk [Chen & Goodman 98]. A kezelésükre számos megoldást dolgoztak ki, melyeket általánosan „simítási technikáknak” nevezünk.

Alapvetően kétféle simítási megközelítés használatos:
1. Visszametszés („backoff”): ilyenkor a rosszul becsülhető N-gram valószínűségeket visszavezetjük a jobban becsülhető (N-1)-gram valószínűségekre [Katz 87]:

$$P(w_K \mid w_{K-1},...,w_{K-(N-1)}) = P_{boN} \cdot P_{ML}(w_K \mid w_{K-1},...,w_{K-(N-2)}) \tag{1.10}$$

ahol a P_{boN} valószínűséget úgy kell megválasztani, hogy a nyelvi modellek által szolgáltatott össz-valószínűség 1 maradjon. Az eljárás az 1 vagy akár a 0-gram szintig folytatható.

2. Interpoláció: a megközelítés lényege, hogy amikor N-gram szinten egy valószínűség rosszul becsülhető – szemben az előző megoldással – nem zárja ki az adott N-gram szintet a valószínűségi becslésből, hanem interpolálja az alacsonyabb rendű értékkel [Jelinek & Mercer 80].

$$P(w_K \mid w_{K-1},...,w_{K-(N-1)}) = \lambda P_{ML}(w_K \mid w_{K-1},...,w_{K-(N-2)}) + (1-\lambda) P_{ML}(w_K \mid w_{K-1},...,w_{K-(N-1)}) \tag{1.11}$$

A simítási eljárások széles választéka áll rendelkezésre, melyek közül a peremfeltételek ismeretében célszerű választani. [6]

1.5.2. Akusztikus modellek

Fonológiai kiejtési modellek

Az alapszintű (fonológiai) kiejtési modelleket hagyományosan kézzel (pl. kiejtési szótár formájában) vagy gépi úton (pl. graféma-fonéma szabályokkal) állítják elő. Amennyiben egy szóhoz több kiejtés is tartozhat, akkor a kényszerített felismerés eszközével – amikor is egyedül csak a helyes szósorozatot engedjük meg a felismerési feladatban – automatikusan is ki lehet választani a beszédadatbázis adott szövegrészletéhez tartozó fonológiai megvalósulást [Young 06]. Így a tanító-adatbázisban az egyes kiejtési variációk valószínűsége a relatív gyakorisággal ML módon becsülhető.

Megjegyzendő, hogy az egyes szavakhoz tartozó kiejtésvariációk valószínűségeinek meghatározására már diszkriminatív tanítási módszereket is alkalmaznak, mely [Schramm & Beyerlein 02] szerint pontosabb felismerési eredményeket biztosít.

[6] Népszerű módszerek: Good-Turing [Good 53], Witten-Bell [Witten & Bell 91], módosított Kneser-Ney [Chen & Goodman 98].

Fonéma akusztikus modellek

A fonéma akusztikus modellek paraméterbecslése alatt az elemi HMM-ek *a* és *b* paramétereinek elemi modellenkénti vagy együttes meghatározását, beállítását értjük. A fő feladat a $b_j(\mathbf{o}_t)$ eloszlások becslése, hiszen a mintaillesztés alapvetően ezek alapján történik.

A HMM állapotokhoz tartozó folytonos valószínűség eloszlásokat (sűrűségfüggvényeket) a legelterjedtebben Gauss függvények szuperpozíciójaként modellezik (12) [Titterington & Smith+ 85] [7]. Ezt a megközelítést a GMM (Gauss Mixture Model) jelzővel szokták illetni.

$$b_j(\mathbf{o}_t) = \sum_{m=1}^{M} c_{jm} G(\mathbf{o}_t, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) \tag{1.12}$$

ahol *M* a Gauss függvények száma, $\boldsymbol{\Sigma}_{jm}$ a kovariancia mátrix $\boldsymbol{\mu}_{jm}$ átlagvektor és c_{jm} súlytényező Gauss függvényenként, tehát

$$\sum_{m=1}^{M} c_{jm} = 1 \tag{1.13}$$

Mivel az o_t jellemzővektorok többé-kevésbé dekorreláltnak tekinthetők, tipikusan diagonális kovariancia mátrixszal szoktak számolni.

Az akusztikus modellparaméterek ML becslése történhet a „k-means" algoritmussal [MacQueen 67], amit az egyes beszédhang modellek *a* és *b* paramétereinek együttes becslésére a Viterbi újraszegmentálással egészítenek ki. A másik gyakran alkalmazott tanítási módszer a Baum-Welch [Baum & Eagon 67] újrabecslő algoritmus futtatása először uniform modellparaméterek esetén, majd a Mixture-splitting eljárással növelt komplexitású modelleken való iteratív alkalmazása [Young 06]. Míg az előbbi esetben szükséges a tanító-adatbázisban a modellezett beszédegységek (hangok, esetleg szavak) határainak kijelölése, a második esetben nem, elegendő csupán a modellszekvencia ismerete. Ezek a módszerek alapvetően az EM (Expectation Maximization) algoritmus [Dempster & Laird+ 77] egyes variánsai.

Az eljárások további részletezését mellőzzük, azok kimerítő leírása megtalálható pl. a [Rabiner & Juang 93] és [Young 06]-ban.

Fontos megjegyezni, hogy az ML akusztikus modell becslését a korszerű rendszerekben gyakorta diszkriminatív tanítás követi. Ennek lényege, hogy a modellek tanító-adatbázishoz való illeszkedési mértékének maximalizálása helyett közvetlenebb módon, pl. MMI (Maximum Mutual Information), MCE (Minimum Classification Error) vagy MPE (Minimum Phone Error) kritériumoknak megfelelően minimalizálja a felismerési hibát [Bahl & Brown+ 86, McDermott 97, Povey & Woodland 02].

[7] GMM helyett alkalmazható más univerzális függvényapproximációs megoldás, pl. mesterséges neurális hálózat is, pl. MLP (Multi Layer Perceptron) [Bourlard & Morgan 93] – ezek ismertetésére külön nem térünk ki.

1.5.3. Adaptáció

Speciális felismerési körülmények és általános modellek esetén a modellparaméterek speciális körülményekhez történő transzformálását nevezzük adaptációnak. Tipikus eset a beszélőadaptáció, amikor rendelkezésre áll egy nagyszámú beszélővel tanított beszélőfüggetlen akusztikai modell, viszont adott esetben tudható, hogy csak egyetlen beszélő hanganyagán kívánunk beszédfelismerést végezni. A beszélőhöz adaptált akusztikus modellekkel a felismerés nem csak lényegesen pontosabb, de egyszersmind gyorsabb is lehet.

Az adaptációnak két alaptípusa lehetséges:

- *Felügyelt adaptáció:* alapvetően „off-line" módon történik a modellparaméterek transzformációja. Ilyenkor szükség van már rögzített és ismert adaptációs tanítóadatra (hanganyag + szövegátirat), mely alapján elvégezzük a (tipikusan lineáris) transzformációt a modelleknek az adaptációs hanganyagra történő jobb illeszkedését célozva.

- *Felügyelet nélküli adaptáció:* ilyenkor nem ismeretes az adaptációs hanganyag tartalma. A felügyelet nélküli adaptáció ezért mindig két fő lépésből áll: felismerés adaptálatlan modellekkel, és így egy közelítő pontosságú szövegátirat előállítása (i); ML lineáris regressziós modelltranszformáció elvégzése az előző pontban előállított átirat segítségével (ii).

Az adaptáció nyelvi, kiejtési, elemi akusztikai szintű is lehet, azonban a továbbiakban csak az elemi akusztikus modellek adaptációjáról fogunk röviden szólni.

Beszélőadaptáció felügyelt MLLR akusztikus modell transzformációval
Talán a legelterjedtebb akusztikus modell adaptációs eljárás a Maximum Likelihood Linear Regression (MLLR) [Leggetter & Woodland 95] módszer, amelyet gépi beszédfelismerésnél alkalmaznak. Ez a ML lineáris transzformáció az elemi akusztikus modellek, azaz a (környezetfüggő) beszédhangmodellek HMM állapotaihoz tartozó folytonos valószínűségsűrűség függvények (GMM-ek) várható érték és (diagonális) kovariancia mátrixait transzformálja oly módon, hogy az adaptációs tanítóadatokon a modellek hasonlósági mértékét valamilyen EM (Expectation Maximization) alapú módszerrel [Dempster & Laird+ 77] maximalizálja.

A GMM várható értékek lineáris transzformálása:

$$\hat{\boldsymbol{\mu}} = W\boldsymbol{\mu} + b \tag{1.14}$$

ahol $\hat{\boldsymbol{\mu}}$ és $\boldsymbol{\mu}$ a transzformált és az eredeti Gauss függvény várható értékek, W a transzformációs mátrix, b pedig az additív tag.

A lineáris transzformáció paraméterei lehetnek globálisak, azaz minden beszédhang GMM paramétereit ugyanúgy módosítjuk. Azonban ez nem mindig praktikus, hiszen a beszédhangok különböző csoportjainál más és más jellegű ejtésmódosulás léphet fel. Ezért az elemi beszédhangmodellek között (ML) lineáris regressziós faépítéssel osztályokat szoktak képezni [Leggetter & Woodland 95], melyeken belül azonosak a transzformációs paraméterek, de csoportonként már különbözőek lehetnek. Az akusztikus modell adaptációs technikák további részletezését mellőzzük, azok megtalálhatók pl. a [Young 06]-ban.

1.6. A felismerési eredmények kiértékelése

Általános beszédfelismerési teszteknél tipikusan bemondások, mondatok sokaságát ismertetjük fel, azaz alakíttatjuk át szöveges szósorozattá. A felismert szavak helyességét emberi munkával előállított referencia-átiratokhoz hasonlítva határozzuk meg bemondásonként. A következőkben röviden ismertetjük a folyamatos beszédfelismerés eredményeinek kiértékelésénél használt metrikákat, illetve a hipotézisvizsgálati módszereket, melyekkel becslést adhatunk az eltérések szignifikanciájára.

1.6.1. Metrikák

Folyamatos felismerésnél többféle metrikát szoktak alkalmazni a beszédfelismerés „jóságának" mérésére. A leggyakrabban használt mennyiség a WER (Word Error Rate), azaz szóhiba arány, amit a következőkben származtatunk.

A felismerési eredményt – ha addig nem olyan formában volt – szósorozattá alakítjuk. A referencia átirathoz a dinamikus programozás módszerével hasonlítjuk, ahol a következő súlyokat rendeljük az egyes lehetőségekhez:

C (helyes, „korrekt" felismerés): 0
S (helyettesítés, „szubsztitúció"): 10
D (törlés, „deletálás"): 7
I (beszúrás, „inzerció"): 7

A kiértékelés alapja a legkisebb összsúlyú összerendelés. A fenti betűjelekkel az adott jelenségek darabszámát jelölve, az alábbi felismerési mérőszámok definiálhatók:

$$Felismerési\ arány\ (Correct\ Rate:"Corr") = \frac{N-S-D}{N}\times 100\%, \tag{1.15}$$

$$Felismerési\ pontosság\ (Accuracy:"Acc") = \frac{N-S-D-I}{N}\times 100\%, \tag{1.16}$$

ahol *N* az összes felismerési egység (szó) száma a referenciaátiratban. A legtöbb alkalmazásnál hibának számít a referenciában nem szereplő szavak beszúrása is, ez csak a felismerési pontosságban jelenik meg. A felismerési pontosság lehet akár negatív is, ha nagy a beszúrások száma.

A felismerési hiba általánosan elfogadott definíciója a következő:

$$Felismerési\ hiba\ (Error\ Rate:"ER") = 100\% - Felismerési\ pontosság = \frac{S+D+I}{N}\times 100\% \tag{1.17}$$

WER: Szó felismerési egységeknél tehát a felismerési hiba a WER. Magyar és egyéb agglutináló nyelvek esetén azonban a szófelismerési hiba bizonyos esetekben túlzottan pesszimista becslést adhat a felismerés jóságáról. (Pl. egy összetett szó két szóként való – egyébként helyes felismerése – egy helyettesítési és egy beszúrási hibát eredményez.)

LER: Egyre inkább elterjedt a LER (Letter Error Rate), azaz a „betű" felismerési hiba, mint metrika használata. Ez a kézi javítás „költségével" az előzőnél jobban korreláló mennyiség. Mi a szóközt is betű értékűnek definiáljuk, egyébként ugyanúgy számoljuk ki karakter egységenként, mint a szóhibaarányt.

A gyakorlatban azonban általában nem a felismerési hiba abszolút értéke a kérdés, hanem leggyakrabban annak megváltozása. Ezen belül is tipikusan a javulás relatív mértéke az érdeklődés tárgya. Ezt az alábbiak szerint definiáljuk mind WER, mind LER esetén.

$$Relatív\ javulás\ (-\Delta ER_{rel}) = \frac{ER_{referencia} - ER_{új}}{ER_{referencia}} \times 100\% \tag{1.18}$$

Végül, gyakorlati szempontból igen lényeges metrika lehet a felismerés időigényének az alakulása is, természetesen adott hardver esetén. Erre az RTF (Real Time Factor) a szokásos mérték.

$$RTF = \frac{felismerésre\ fordított\ idő}{felismert\ beszéd\ hossza} \tag{1.19}$$

Tehát az alacsonyabb értékek a jobbak.

1.6.2. Szignifikancia-vizsgálatok

Attól, hogy az egyik felismerési teszt során jobb eredményt kaptunk, mint a másikban, még nem jelenthetjük ki 100% biztonsággal, az utóbbi megközelítés általánosságban véve is jobb, hiszen véges méretű teszthalmazzal dolgozhatunk csak.

Hasznos lehet a felismerési hiba relatív csökkenése és a hasonló mérőszámok mellett a szignifikanciaszintet is megadni, ami megmutatja, hogy mekkora a tévedés valószínűsége a tekintetben, hogy az eredmények alapján jobbnak minősítettük az egyik megközelítést a másiknál.

Például, a felismerési arányokon alapuló *2 mintás Z-próbával* egy valószínűségi becslést kaphatunk arról, hogy két eltérő, de ugyanolyan körülmények között tesztelt felismerési megközelítés közül az egyiket a másiknál jobbnak ítélve, mennyire lehetünk biztosak abban, hogy jól döntöttünk.

Legyen p_1 annak a becsült valószínűsége, hogy az első felismerő rendszerrel egy szót helyesen ismerünk fel, p_2 pedig a második felismerő rendszerre vonatkoztatva jelentse ugyanezt. Vagyis, a p_1 és p_2 értékeket jól közelíti a két felismerő rendszernél mért szó felismerési pontosság.

n független felismerési kísérlet esetén az alábbi módon számított *Z* eloszlása standard normálisak tekinthető.

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 \cdot (1 - p_1)}{n} + \frac{p_2 \cdot (1 - p_2)}{n}}} \tag{1.20}$$

Adott p_1, p_2 felismerési pontosságok és *n* tesztfelvétel esetén, $1 - \Phi(Z)$ megadja annak a döntési hibának a közelítő valószínűségét mely szerint az első rendszer jobb, mint a második.

Az eljárás gyengéje, hogy csak akkor ad megalapozott becslést, ha a felismert szavak egymástól függetlenek, illetve a fenti definíció szerinti Z valóban jól közelíti a normális

eloszlást. Míg izolált szavas teszteknél a szóhibaarányra vonatkozóan megfelelően nagy mintaszámok esetén ez teljesül is, folyamatos beszéd felismerésekor a szó- és betűhibaarányra vonatkoztatva már kevésbé (hiszen a nyelvi és kiejtési modellezésnél pont abból indulunk ki, hogy az egymás utáni egységek nem függetlenek egymástól).

E problémák miatt – főként a folyamatos beszédfelismerési eredmények szignifikancia-vizsgálatához – a fentinél összetettebb módszereket szoktak használni. Ilyen például a NIST (National Institute of Standards and Technology) ajánlásban szereplő nem parametrikus *Wilcoxon előjeles rang teszt* [Wilcoxon 45]. A módszer alkalmas arra, hogy azonos tesztadatokon futtatott A és B felismerő rendszer szegmenspárokon összehasonlított eredményei alapján becsülje meg annak biztonságát, hogy B jobb, mint A. Ehhez az adott szegmenseken mért WER (vagy LER) értékeket összehasonlítja, a különbségeket rangsorolja, majd a javulás/romlás szerint előjelezi.

Legyen W^+ a pozitív rangok összege, n pedig az (egymástól függetlennek tekintett) szegmensek száma. N > 8 esetén W^+ normális eloszlásúnak vehető, melynek paraméterei:

$$\mu = \frac{n(n+1)}{4} \tag{1.21}$$

$$\sigma^2 = \frac{n(n+1)(2n+1)}{24} \tag{1.22}$$

Ekkor a teszt statisztika megközelítőleg normál eloszlású:

$$Z = \frac{W^+ - \mu}{\sigma} \tag{1.23}$$

Innentől a hipotézisvizsgálatoknál megszokott módon becsülhetjük az elsőfajú döntési hiba valószínűszínűségét [Kanji 94], Daniel [78].

Az értekezésben – ahol nem jelöljük másként – egyoldalas $p = 0.05$ szignifikancia szint alatt (*0.95* konfidencia szint felett) fogadjuk el a nullhipotézist, miszerint a *B* felismerő jobb, mint az *A*. A szegmensek, melyek felismerési részeredményeit rangsorolva kaptuk a javulás szignifikanciájára vonatkozó eredményeket általában mondatok vagy nagyobb közlési egységek.

A szignifikáns relatív javulásokat, illetve ahol csak abszolút felismerési számokat közlünk, a referenciától szignifikánsan jobb eredményeket dőlt betűvel jelenítjük meg.

2. Tudásforrás-integráció a WFST keretrendszerben

Mint az előző fejezetben bemutattuk, a beszéd beszélőfüggetlen modellezésére jól használhatóak a rejtett Markov-modellek. A rejtett Markov-modell felismerési hálózatok 1.4-ben említett minimális kiterjesztésével a beszédfelismerési feladatok széles skálája kezelhető. Nem ejtettünk viszont szót arról, hogy a különféle tudásforrásokat (úgymint nyelvi, kiejtési, elemi akusztikai stb. modellek), hogyan építsük össze egyetlen HMM felismerési hálózattá. Az egyszerűbb nyelvi (bigram), kiejtési (változatok nélküli) és fonéma akusztikus modell tudásforrások integrációja nem okoz problémát, amíg a kiejtési modell környezetfüggetlen (monofón) beszédhangmodellekre épül. Azonban különösen folyamatos beszédfelismerésnél a magasabb rendű (N=3, 4, …) nyelvi modellek, illetve a környezetfüggő beszédhangmodellek megfelelő („cross-word") integrálása, valamint a végeredményül előálló hálózat hatékony optimalizációja már korántsem triviális feladat.

Sokáig a speciális tudásforrások felismerő rendszerbe építéséhez és az eredményül előálló HMM-hálózat optimalizációjához speciális célalgoritmusok használata volt általános [Steinbiss & Tran+ 94], [Odell & Valtchev+ 94], [Ortmanns & Ney+ 96]. Azonban ezek a megközelítések nem teljesen rugalmasak, kötöttek lehetnek bizonyos modellstruktúrák (pl. csak bigram nyelvi modell, csak szóbelsejei trifón modellezés stb.). Továbbá, az optimalizálás jobbára modellszintenként történik szuboptimális felismerési hálózatot eredményezve. Így akár új tudásforrás hozzáadása, akár globális optimalizáció a rendszer jelentős továbbfejlesztését teheti szükségessé.

A látszólag jelentősen különböző tudásforrásokról (úgymint az N-gram nyelvi modell, kiejtési modellek stb.) ugyanakkor már pár évtizede ismert, hogy leírhatók (súlyozott) véges állapotú gépekkel. Néhány éve az AT&T kutatói megmutatták, hogy a véges állapotú gépek súlyokkal és kimeneti szimbólumokkal történő kiterjesztésével a beszédfelismerésnél alkalmazott tudásforrások modellezhetők, mi több, a tudásforrások kombinálása, és az eredmény globális optimalizálása matematikailag jól definiált műveletekkel hatékonyan megoldható [Mohri 97], [Mohri & Pereira+ 02]. A keretrendszert a WFST-k (Weighted Finite State Transducer), azaz súlyozott véges állapotú átalakítók után nevezték el.

A témával kapcsolatban bőséges szakirodalom áll rendelkezésre – beleértve a magyar nyelvű beszédfelismerést is [Szarvas & Furui 03], [Szarvas 03] – ezért hangsúlyozottan a teljesség igénye nélkül foglaljuk össze a WFST-kel kapcsolatos alapismereteket és beszédfelismerési alkalmazásokat.

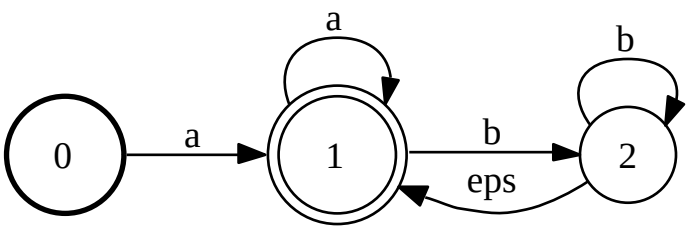
2.1. A WFST modellezés alapjai

A következőkben röviden bemutatjuk, mely általánosításokkal juthatunk a „hagyományos" véges állapotú gépektől a súlyozott véges átalakítókig, valamint áttekintjük a legfontosabb műveleteket.

2.1.1. A WFST származtatása

Véges állapotú gép – FSA: A véges állapotú gépeket (FSA: Finite State Acceptor) olyan (félgyűrű felett értelmezett) matematikai objektumoknak is tekinthetjük, melyek egy adott szimbólumsorozathoz egy bináris értéket rendelnek, attól függően, hogy az automata elfogadja-e a szimbólumsorozatot vagy sem.

Egy véges automata állapotokkal, irányított, szimbólummal ellátott átmenetekkel reprezentálható. A grafikus megjelenítési konvenciók szerint a kezdőállapotot vastag vonallal, a normál állapotokat normál vonallal, a végállapotot pedig kettős vonallal rajzolt körök jelölik.

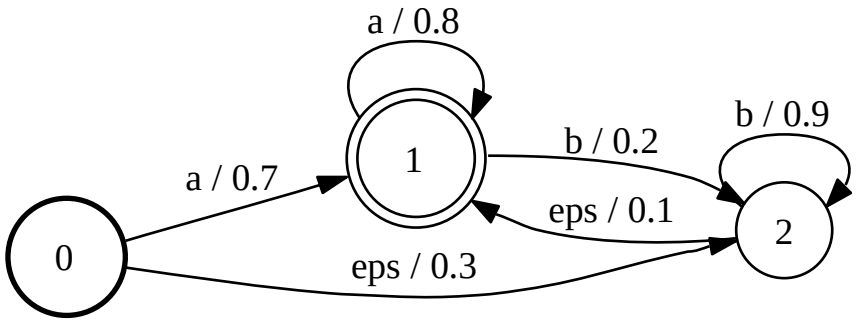


2.1. ábra. Illusztráció: Véges állapotú elfogadó automata (FSA: Finite State Acceptor).

Az állapotátmenetek a bejövő szimbólumok hatására jöhetnek létre, kivéve az ún. epszilon (eps) él esetén, melyen a tranzíció e nélkül is létrejöhet.

A fenti automata (2.1. ábra) tehát elfogadja az „abba” szimbólumsorozatot, de nem fogadja el a „baba” szimbólumsorozatot, tehát első esetben 1, a másodikban 0 értéket rendelhet hozzájuk.

Súlyozott véges állapotú gép – WFSA: A véges állapotú elfogadó automaták adott szimbólumsorozathoz tehát csak bináris leképezés mentén képesek kimeneti értéket rendelni. Ez azonban nem mindig praktikus, hiszen a világ nem midig fekete-fehér, gyakran szürke is lehet (nem mindegy azonban, hogy milyen mértékben…). Ha a véges állapotú gépek állapotátmeneteit súlyokkal látjuk el, akkor az automata a megfigyelt szimbólumsorozathoz már számszerű értéket, pl. valószínűséget, ill. súlyokat rendelhet (WFSA: Weighted Finite State Acceptor). A sztochasztikus, valószínűségeket társítani képes automaták ekvivalensek a Markov-modellekkel (Markov láncokkal), és hasonlóan sok célra használhatók, pl. nyelvi modellezésre is. Ezek az automaták a valószínűségi félgyűrű felett értelmezettek. A valószínűségeket, illetve súlyokat a bemenő szimbólum után „/” jellel elválasztva ábrázoljuk, ha eltérnek az alapértelmezettől, lásd 2.2. ábra.

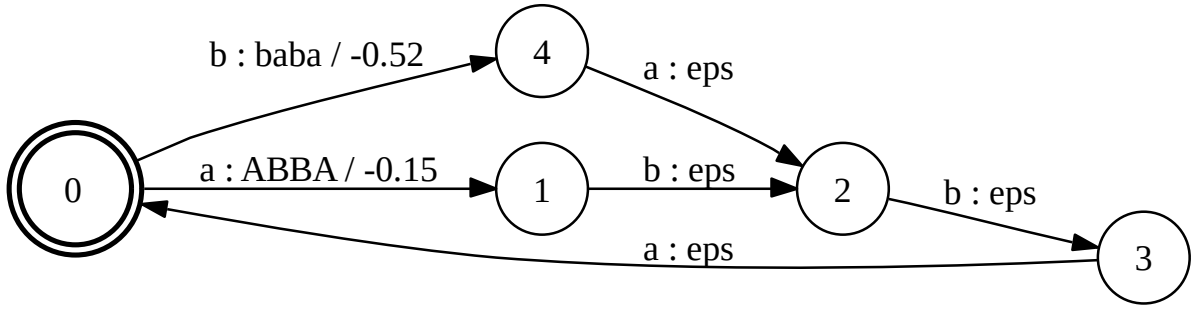


2.2. ábra. Illusztráció: Súlyozott véges állapotú elfogadó automata (WFSA: Finite State Acceptor), valószínűségi súlyozással.

Numerikus stabilitási problémák miatt a gyakorlatban azonban nem valószínűségekkel, hanem azoknak negatív logaritmusával szoktak számolni. Ekkor a legkisebb összsúly adja meg egy szimbólumsorozathoz rendelődő mennyiséget. Az ilyen véges átalakítókat az ún. tropikus félgyűrű felett értelmezik, pl. 2.3. ábra [Mohri & Pereira+ 02].

Súlyozott véges állapotú átalakító – WFST: A súlyozott véges állapotú gépek tehát képesek *súlyt* rendelni egy adott bejövő szimbólumsorozathoz. Azonban a beszédfelismerésnél és sok más alkalmazásnál nem közvetlenül e számértékre van szükség, hanem a felismerés eredményére, azaz a legkisebb összsúlyú úthoz tartozó kimenő szó (szimbólum) sorozatra.

A véges állapotú gépek következő általánosítása, mely a fenti problémára megoldást jelent, az a szimbólumok lecserélése bejövő-kimenő szimbólum *párokra*. Ezeket az objektumokat súlyozott véges állapotú átalakítóknak nevezzük, mert a bejövő szimbólumsorozatot egy kimenő szimbólumsorozatra képzik le, miközben a leképzéshez egy súlyt is társítanak (WFST: Weighted Finite State Transducer). Ahogy a 2.3. ábrán látható, a bemenő szimbólum után „:”-tal elválasztva tüntetjük fel a kimenő szimbólumot. Az „eps” kimenő szimbólum azt jelenti, hogy az adott átmenetnél nem írunk ki semmit.



2.3. ábra. Illusztráció: Súlyozott véges állapotú átalakító automata (WFST: Finite State Transducer), logaritmikus valószínűségi súlyozással

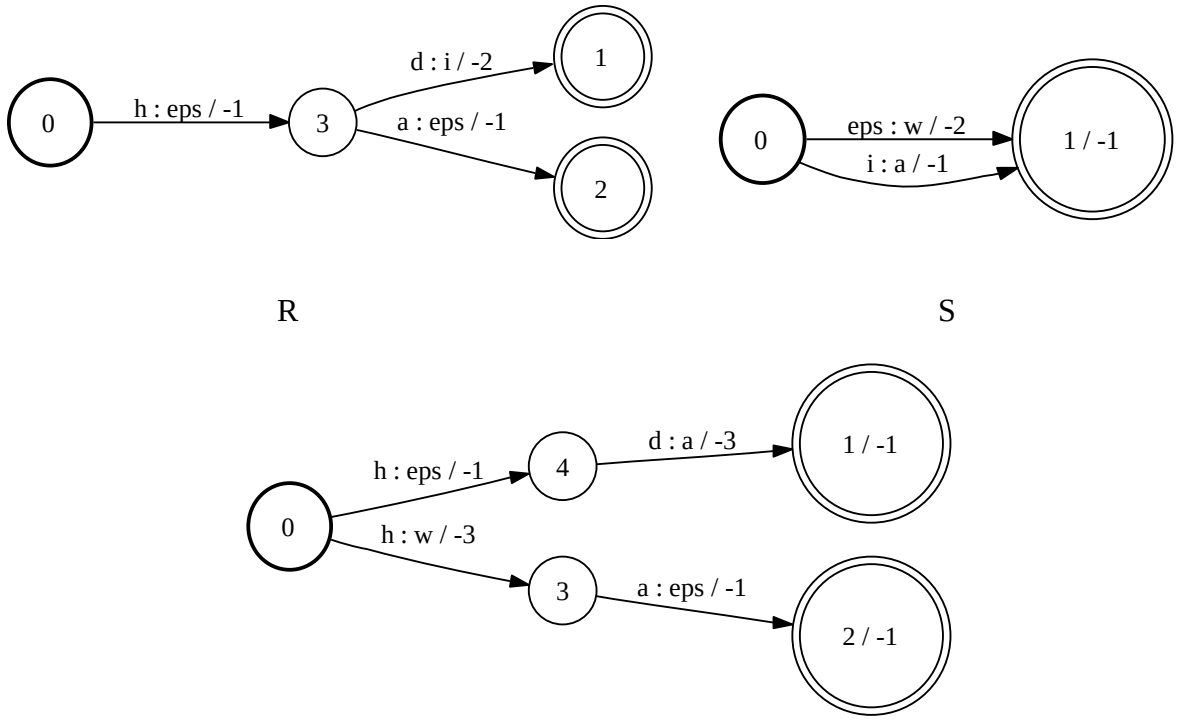
Azonos bejövő-kimenő szimbólumok esetén az átalakító nem alakít át semmit, így kaphatunk súlyozott véges állapotú gépet. Ha még az adott (tropikus félgyűrű feletti) esetben a súlyokat is 0-nak választjuk, akkor a „hagyományos” véges állapotú gépekhez jutunk vissza.

2.1.2. Műveletek súlyozott véges átalakítókkal

A súlyozott véges állapotú átalakítókon (továbbiakban csak átalakítók, vagy WFST-k) végezhető műveletek általában a véges automatákon értelmezett műveletek egyszerű általánosításai. Ilyenek például az *unió, konkatenáció, Kleene-lezárás* [Mohri & Pereira+ 02]. A következőkben röviden összefoglaljuk a beszédfelismerési alkalmazásokban kulcsfontosságú egyéb műveleteket.

Kompozíció: Mint korábban említettük, a WFST-k valamely bejövő szimbólumsorozatot egy kimenő szimbólumsorozatra (és súlyra) képeznek le. A kimenő szimbólumsorozat viszont gyakran egy másik átalakító bemenete, melyet a második átalakító egy újabb kimenő szimbólumsorozatra képez le. Ilyenkor természetes igény lehet, hogy a két átalakító helyett eggyel oldjuk meg a feladatot.

Legyen R átalakító, mely az U bejövő szimbólumsorozatot a V szimbólumsorozatra képzi le, S pedig az az átalakító, mely a V-t a W-ra képzi le. Ekkor T=R o S a két átalakító kompozíciója, az U-ról W-re történő leképzést valósítja meg a két különálló leképezés összsúlyával, lásd 2.4. ábra.

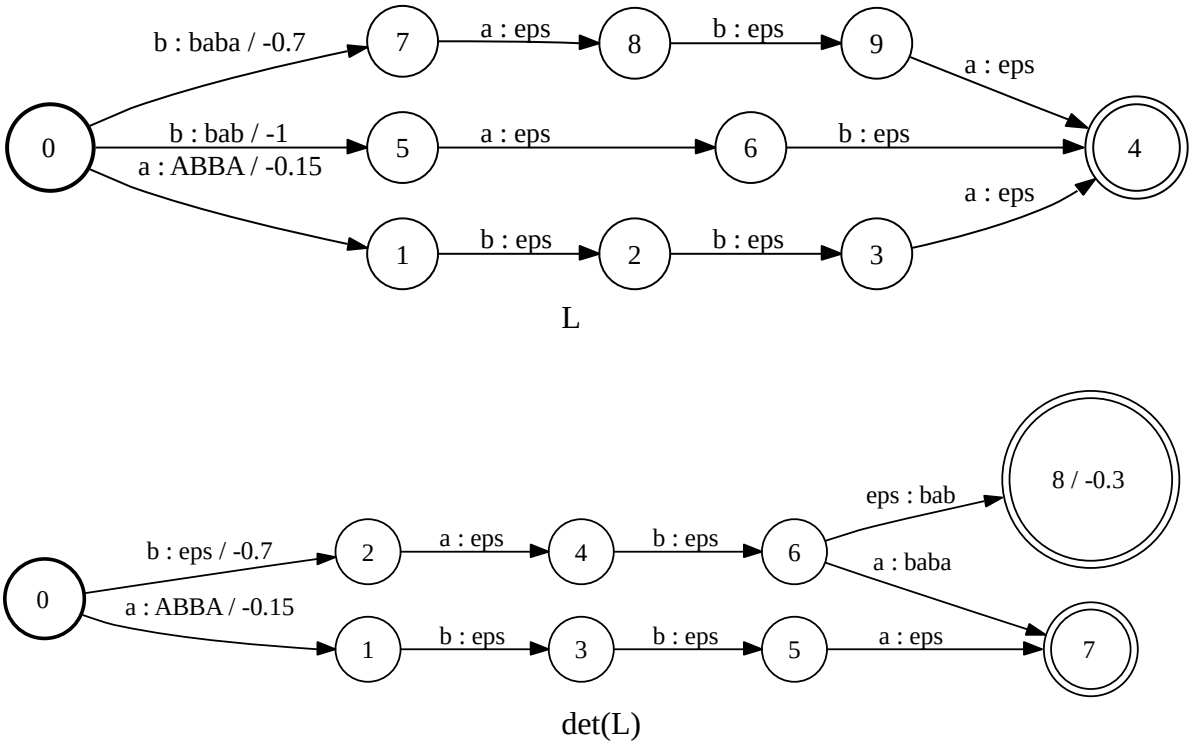


2.4. ábra. A súlyozott véges állapotú átalakítók kompozíciójának szemléltetése

Determinizáció: Hasonlóan a véges állapotú gépekhez, a WFST-k is lehetnek determinisztikusak, vagy nem-determinisztikusak. Az előbbi esetben bármely adott állapot és bejövő szimbólum esetén pontosan egy következő állapot lehetséges, míg utóbbi esetben ez nem áll fenn.

Bizonyos nem-determinisztikus hálózatok ekvivalens determinisztikus véges átalakítókká alakíthatók. Ezt az átalakítást nevezzük determinizációnak. Ekvivalensnek akkor tekintünk két átalakítót, ha bármely adott bejövő szimbólumsorozathoz ugyanazt a kimenő szimbólumsorozatot és ugyanazt a súlyt rendelik.

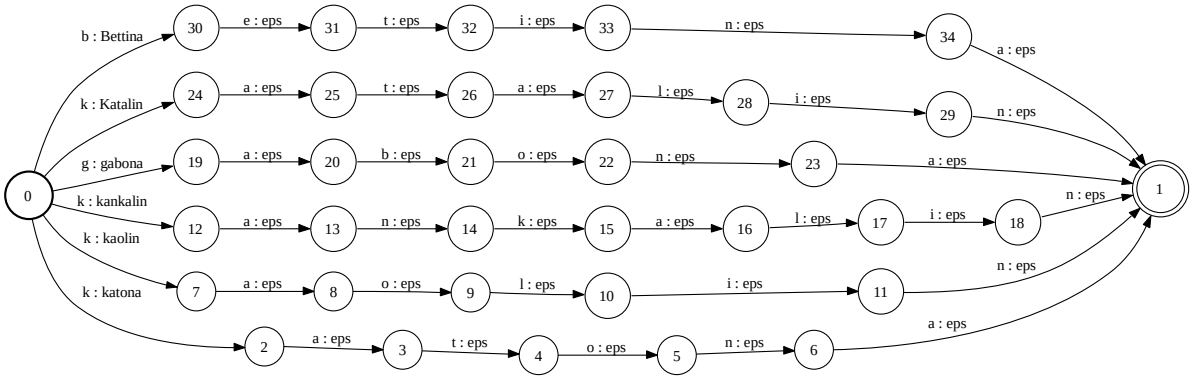
A determinisztikus átalakítók alapvető előnye a nem-determinisztikusakkal szemben, hogy irredundánsak számítási szempontból, hiszen minden új bejövő szimbólum esetén csak egy továbblépést kell kiértékelni. A súlyozott átalakítók viszont – ellentétben a súlyozatlan automatákkal – nem minden esetben determinizálhatók. Szerencsére a beszédfelismerési alkalmazásokban a legtöbb átalakító determinizálható, vagy egyszerű transzformációkkal azzá alakítható. Mindenesetre, minden aciklikus súlyozott átalakítható determinizálható. Bővebb információk az eljárásról a [Allauzen & Mohri 02]-ban találhatók.



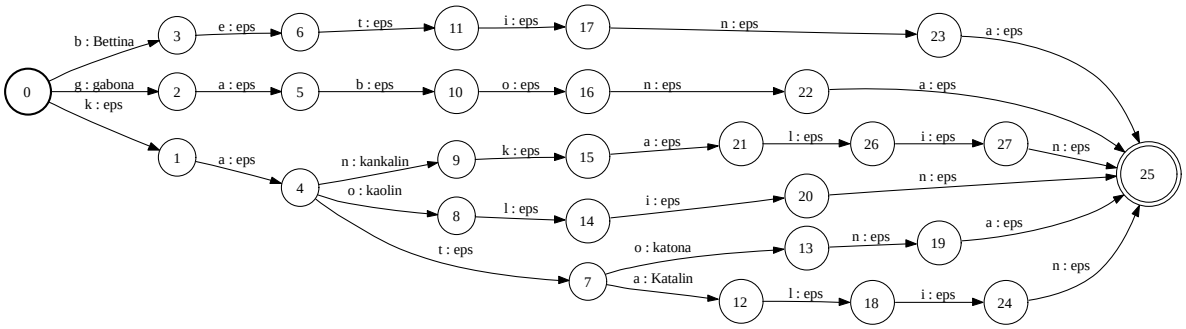
2.5. ábra. A súlyozott véges állapotú átalakítók determinizációjának szemléltetése

Minimalizáció: Az, hogy egy súlyozott véges átalakító determinisztikus, nem jelenti azt, hogy a lehető legkevesebb állapottal is valósítja meg az adott leképezést. A „hagyományos” véges állapotú gépek minimalizációjára számos klasszikus eljárás használható [Aho & Hopcroft+ 74] [Bauer 88]. Hasonlóan a véges állapotú gépekhez, minden determinisztikus súlyozott véges állapotú gép minimalizálható [Mohri 97].

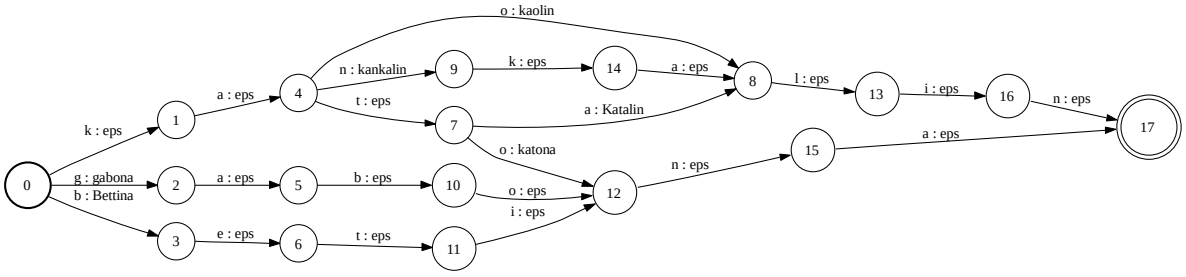
A WFST-k esetében a súlyok és a kimeneti címkék megfelelő kezelésével visszavezethetjük a minimalizációs problémát a hagyományos automatákéra. Az első lépés a súlyok és kimenő szimbólumok összevonása speciális „súlyokká”, így az átalakító helyett súlyozott véges automatát (WFSA-t) kapunk. A következő feladat a súlyok előretolása oly módon, hogy az átalakító ekvivalens maradjon. Erre a [Mohri & Riley 01]-ban részletezett algoritmus használható. Ezután az automata már klasszikus módszerekkel minimalizálható.



D



det(D)



min(D)

2.6. ábra. A súlyozott véges állapotú átalakítók optimalizációjának, illetve a determinizált és minimalizált alak különbségének szemléltetése.

2.2. Beszédfelismerési hálózatépítés

A következőkben röviden vázoljuk a WFST-k beszédfelismerési hálózatok előállításában játszott szerepét.

2.2.1. Tudásforrás-integráció

A beszédfelismerésnél használt, kimeneti címkékkel kiegészített rejtett Markov-modell felismerési hálózatok strukturálisan ekvivalensek a súlyozott véges állapotú átalakítókkal. Azaz, az elemi rejtett Markov-modell állapotok folyamatos valószínűségsűrűség-függvényeit leszámítva, bármely beszédfelismerési hálózat WFST-ként leírható. (Hasonlóan [Ljolje & Riley+ 99]-hoz, az egyszerűség kedvéért az elemi HMM-ek helybenmaradási és továbblépési valószínűségeit sem a WFST hálózatban rögzítjük.) Azonban nemcsak a teljes felismerési hálózat reprezentálható WFST formátumban, hanem az egyes tudásforrások is, úgymint a nyelvi, kiejtési, elemi akusztikus modell struktúrák, környezetfüggőségi stb. modellek.

A WFST keretrendszer óriási előnye, hogy bármilyen WFST-be kódolható tudásforrás rendkívül egyszerűen, a kompozíció elemi műveletével integrálható a felismerési hálózatba. Hiszen két tudásforrás kompozíciója csak olyan szimbólumsorozatokat enged meg, amelyek mindkét tudásforrás automatája megenged. Tropikus félgyűrű felett értelmezett WFST pedig a súlyt is megfelelően (minimális útvonalsúly) rendeli a kimenő szimbólumsorozat mellett a kimenethez.

Így a különböző modellszintek ötvözése – a felismerési kaszkád – elegánsan, hatékonyan implementálható matematikai műveletekkel állítható össze. Ehhez csak az egyes tudásforrások súlyozott véges állapotú átalakítóvá történő konverziója szükséges.

Egy általános folyamatos beszédfelismerési modell-hierarchiából a következőképpen állítható össze felismerési hálózat.

Legyen

- H az elemi akusztikus szintről környezetfüggő beszédhangszintre leképező modell
- C a trifónról fonémára leképező környezetfüggőségi modellek,
- L a fonológiai szintről szószintre leképező („szótár”) modell,
- G pedig a nyelvi modell

WFST reprezentánsa. Ekkor a felismerési WFST hálózat az alábbi kompozíciósorozattal adódik.

$$\textit{Felismerési hálózat} = \mathrm{H} \circ \mathrm{C} \circ \mathrm{L} \circ \mathrm{G} \tag{2.1}$$

ahol nem tüntettük fel az optimalizációt, mely történhet determinizációval vagy minimalizációval vagy esetleg el is maradhat.

Ekkor a beszédfelismerési feladat a következő:

$$\hat{W} = \textit{legjobb\_útvonal}\left(\textit{Felismerési hálózat} \mid O\right) \tag{2.2}$$

amit a WFST felismerési hálózat triviális HMM-é konvertálásával (a GMM-ek megfelelő élekhez társításával, majd él- csomópont konverzióval) készíthetünk elő, és az 1.4 szerinti Viterbi algoritmus-kiterjesztésekkel oldhatunk meg.

2.2.2. Tudásforrások WFST formátumra konvertálása

Az egyes tudásforrásokat az integráció előtt WFST formára kell hozni. A következőkben röviden áttekintjük az egyes tudásforrások súlyozott véges állapotú átalakítóvá történő konverziójának lépéseit.

Nyelvi modell (G): Gyakorlatilag a szavak kapcsolatait leíró gráf már triviálisan alakítható WFST formátumra, a bejövő és kimenő szimbólumoknál is ugyanazt a szót kell feltüntetni. Az egyes szavak közti átmenetekhez rendelhetünk a valószínűségeknek megfelelő súlyokat.

Az N-gram modellek WFST formába öntése a gyakorlatban általában közelítő módszerrel történik, ugyanis a visszametszéses (back-off) simításnál használ N-1 gram-ra történő átmenet a legkönnyebben párhuzamosan valósítható meg a „korrekt” N-gram ággal (2.7. ábra). Vagyis, ha egy adott szó N-es esetén az N-gram és az N-1 gram valószínűség is adott a visszametszéses (back-off) ágon, akkor az eredeti specifikációval szemben a felismerés során a nagyobb valószínűségű útvonal lesz kiválasztva, nem pedig konstans módon az N-gram útvonal.

Szemléltetésül tekintsük az alábbi, végletesen egyszerű, „ARPA” formátumú bigram nyelvi modell megadást.

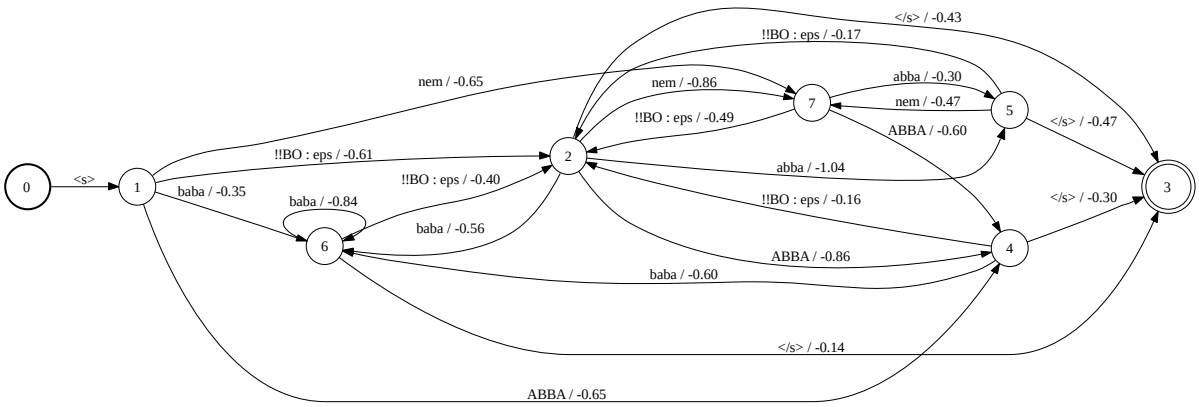
```
\data\
ngram 1=6
ngram 2=11

\1-grams:
-0.4393327   </s>
-99    <s>    -0.61182
-0.8653014   ABBA   -0.1627273
-1.041393    abba   -0.1760913
-0.5642715   baba   -0.4057654
-0.8653014   nem    -0.4900862

\2-grams:
-0.6532125   <s> ABBA
-0.3521825   <s> baba
-0.6532125   <s> nem
-0.30103     ABBA </s>
-0.60206     ABBA baba
-0.4771213   abba </s>
-0.4771213   abba nem
-0.146128    baba </s>
-0.845098    baba baba
-0.60206     nem ABBA
-0.30103     nem abba

\end\
```

Ennek WFST alakja a 2.7. ábrán látható.



2.7. ábra. Visszametszéses „backoff” 2-gram nyelvi modell súlyozott véges állapotú átalakító (WFST) alakjának szemléltetése.

Szótármodell (L): A szósorozat–fonémasorozat leképzés tipikusan szótár jellegű kiejtési modellek alapján történik. A szótár lehet kézi vagy gépi úton előállított, az általános formátum a következő, egymás alatti bejegyzésekből áll:

```
szóalak fonémasorozat [log valószínűség]
```

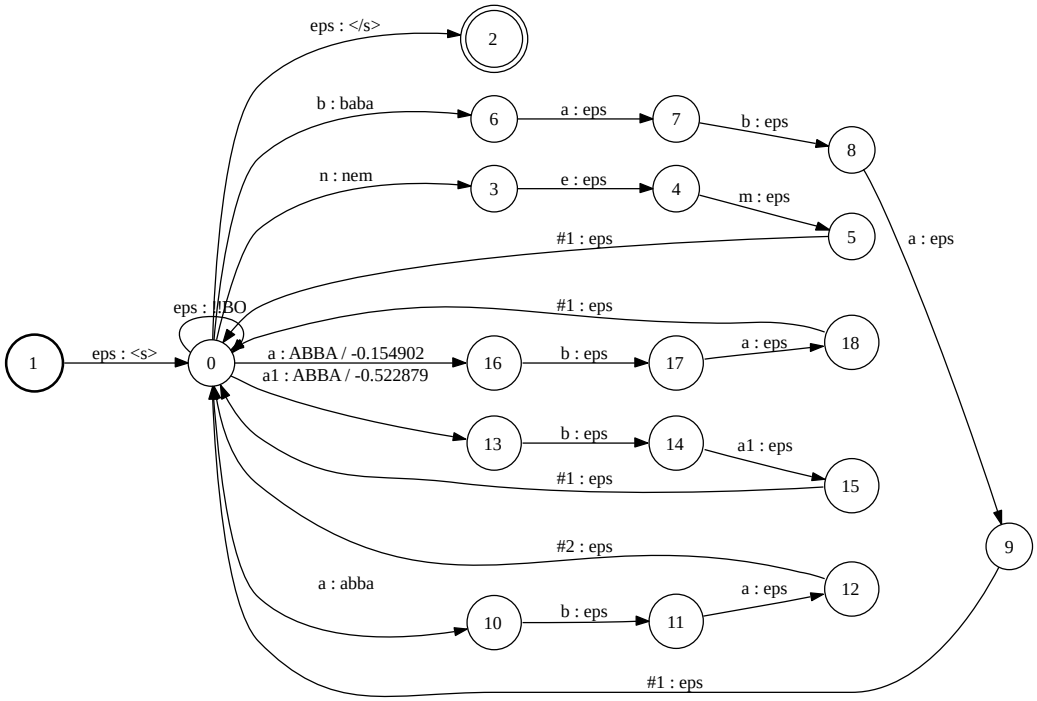
Egy szóalakhoz több kiejtés is tartozhat. A szótár-WFST képzése bejegyzésenként történhet a következő módon: a kezdő állapotból szekvenciálisan következnek az adott kiejtéshez tartozó fonémák, az elsőhöz a kiejtés (log) valószínűségét társítjuk súlyként, és a szóalakot kimenő szimbólumként. A többi fonémához epszilon kimeneti szimbólumot és 0 súlyt rendelünk. Az utolsó fonéma után elfogadó (vég-) állapot következik, ami folyamatos felismerésnél egy központi állapothoz való visszahurkolást jelent. (Különben nem tudnánk több szót felismertetni egymás után.)

Ha determinizálni akarjuk a szótár-WFST-t – ami pedig ajánlatos –, célszerű az azonos fonémasorozatú (homofón) szavak átirata esetén az utolsó állapot előtt egy speciális, sorszámozott szimbólumot (pl. #1, #2 stb.) beilleszteni, amit a későbbiek során eltávolítunk a hálózatból [Mohri & Pereira+ 02].

A szótár transzducer képzését az alábbi példával szemléltetjük:

```
ABBA          á b á  log(0.3)
ABBA          a b a  log(0.7)
abba          a b a
baba          b a b a
nem           n e m
```

mint látható, a rövid és hosszú mássalhangzókat nem különböztetjük meg. A WFST alak a 2.8. ábrán látható.

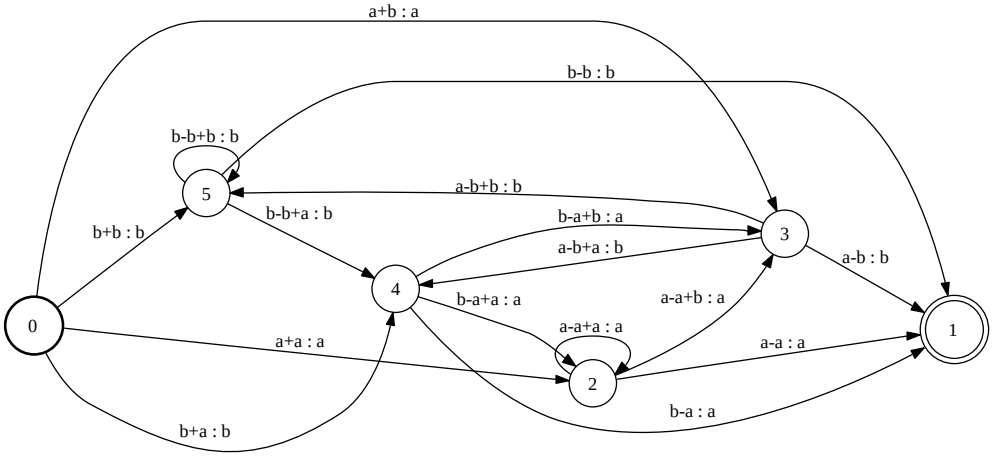


2.8. ábra. Folyamatos beszédfelismeréshez készült szótár modell súlyozott véges állapotú átalakító (WFST) alakjának szemléltetése.

Általános trifón környezetfüggőségi modell (C): Az általánosított trifónok esetén minden fonológiai környezetet megkülönböztetünk. A feladat tehát az egyes fonémákat a fonológiai környezetüknek megfelelő trifónokra leképezni. Ennek véges állapotú átalakítóval történő megvalósítása egyszerűen az összes hangkapcsolati lehetőség szisztematikus lekezelésével megoldható. Természetesen az ilyen hálózat mérete jelentős lehet.

Alább bemutatunk egy lehetőséget környezetfüggőségi átalakító képzésére kételemű ABC (fonéma alaphalmaz) esetén, melyből már a több elemű eset általánosítható.

Legyen a fonológiai ABC a következő: *a,b*. Az általánosított bi- és trifónokat fonémákra képző véges átalakítót a 2.9. ábrán vázoljuk.



2.9. ábra. A környezetfüggőségi transzducer szemléltetése kételemű ABC esetén.

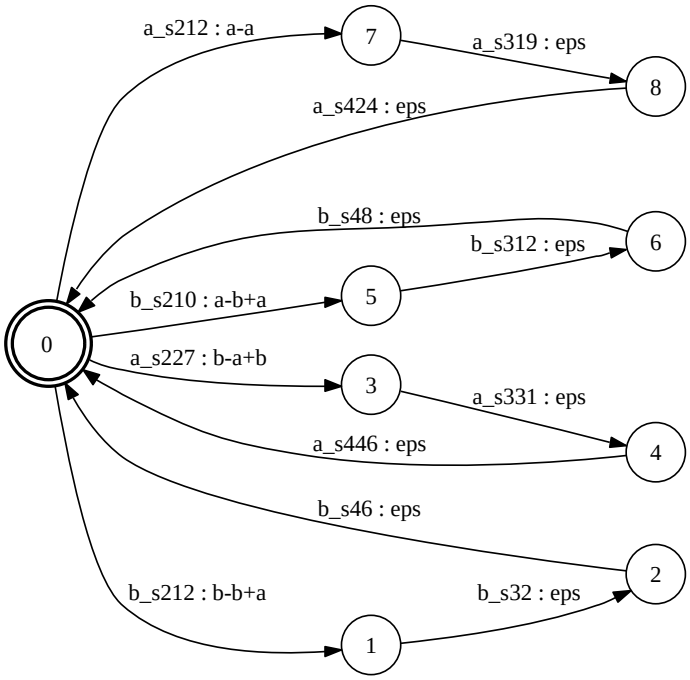
Trifón kiejtési modell (H): A trifón kiejtési modell feladata az elemi akusztikus modelleket megtestesítő rejtett Markov-modell állapotok általánosított trifónokra történő leképezése. A különféle trifón állapotcsoportosítási eljárásokkal maga a leképezés realizálható pl. fonémánkénti döntési fa formájában, az aktuális teendőnk tehát annyi, hogy ezt a leképezést WFST formára hozzuk.

Az első lépés a trifón-HMM állapot (legelemibb akusztikus modell) reláció szótár formátumra hozása. Ezt úgy tehetjük meg, hogy az összes elvi általános trifón leképezését meghatározzuk, és egy szótár állományban felsoroljuk. Tulajdonképpen hasonlóan járunk el, mint a szótármodellnél, csak itt alacsonyabb szinten, az egyes hangokat hangrészletekre képezzük le. Ez a szótár tehát annyi bejegyzést tartalmaz, ahány elvi általános trifón lehetséges, azaz P^3 számút, ahol P ~ 40 a fonológiai kategóriák száma.

Szemléltetésként tekintsük egy részletét egy lehetséges trifón kiejtési szótárnak:

```
a-a    a_s212 a_s319 a_s424
b-a+b  a_s227 a_s331 a_s446
a-b+a  b_s210 b_s312 b_s48
b-b+a  b_s212 b_s32 b_s46
```

Innen hasonlóan folytathatjuk, mint az L Szótármodell esetében. A 2.10. ábra a fenti környezetfüggő beszédhangmodell struktúra WFST-megfelelőjét ábrázolja.



2.10. ábra. Az elemi akusztikus (HMM állapot-) szintről környezetfüggő beszédhangmodellekre leképző transzducer részlet szemléltetése.

2.2.3. Optimalizálás

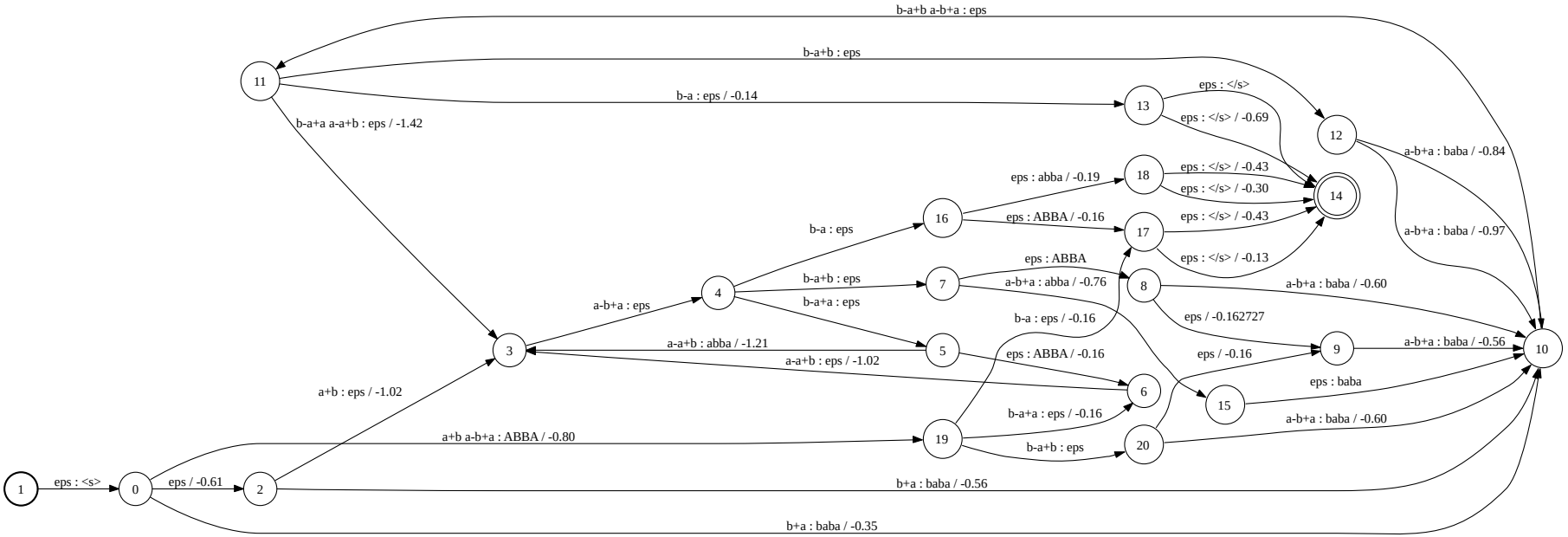
Elméletileg a beszédfelismerési tudásforrások kompozíciójával előállított beszédfelismerési hálózat teljes mértékben alkalmas a mintaillesztésre. Gyakorlati szempontok alapján mégsem ily módon alkalmazzák, mert a folyamatos beszédfelismerés számításigénye a mintaillesztési folyamatban igen jelentős lehet, ezért a valós idejű beszédfelismeréshez általában optimalizációs technikák alkalmazása szükséges.

A felismerési hálózat determinizálása tulajdonképpen a felismerési szótár fastruktúrába fejtésének az általánosításaként is felfogható. Hiszen a determinizált szótár (L) ekvivalens az (elölről) fába fejtett szótárral. Ha viszont az egész felismerési hálózaton végezzük a determinizációt, az magasabb szintű optimalizációt eredményez.

A determinizált hálózat természetesen nem feltétlenül optimális a memóriaigény szempontjából, ilyenkor a minimalizálás alkalmazása lehet célravezető.

Továbbá nemcsak veszteségmentes optimalizálást végezhetünk, hanem a kis valószínűségű felismerési útvonalak (off-line) levágásával csökkenthetjük a hálózat méretét, amivel az on-line keresést is gyorsíthatjuk (vagy inkább a memóriaigényt csökkenthetjük) némi felismerési hibanövekedés árán.

A 2.11. ábrán – az ábrázolhatóság végett – a H átalakító nélküli integrált, optimalizált WFST felismerési hálózatot mutatjuk be. Az integráció során a #1, #2, … szimbólumokat epszilon szimbólumokra cseréltük. Amint az ábrán észrevehető, csak az *a* és *b* fonémákat tartalmazó kiejtések maradtak meg, mivel a C modell csak ezeket tartalmazta.



2.11. ábra. Az integrált és optimalizált C o L o G WFST szintű felismerési hálózat szemléltetése.

3. Célkitűzések

Az szerző általános célja a magyar nyelvű beszéd minél pontosabb, de kézben tartható számításigényű gépi felismerése[8]. Az értekezésben az elsődleges cél a *magyar nyelvi jellegzetességekkel kapcsolatos modellezési kérdések* megválaszolása.

Konkrét, *tézisekben is megjelenő célkitűzések* a magyar nyelvű gépi beszédfelismerés témakörében a következők voltak:

I. A magyar nyelvű *fonetikai koartikuláció* modellezésének vizsgálata az általános gépi beszédfelismerés szempontjából. Azaz, a tárgy a beszédhangok egymásra hatásának vizsgálata, a modellezés miként je.
II. A magyar nyelvre jellemző *fonológiai koartikuláció* modellezésének vizsgálata általános gépi beszédfelismeréshez. Másképpen fogalmazva, a hasonulási, egybeolvadási és egyes hangkiejtési szabályok alkalmazásának módja, szükségessége merül fel kérdésként.
III. A magyar nyelv *lexikai modellezésének* vizsgálata spontán nyelvű beszéd felismeréshez. Itt a nyelvünk morfológiai gazdagsága okozta kihívásokra (nagyszámú szóalak, ritka szóalakok nagy mennyisége, szótáron kívüli szavak magas aránya) adható válasz keresése a feladat alkalmas lexikai egységek megválasztásával (nyelvi, statisztikai morfémaszerű egységek).
IV. A magyar nyelv *kiejtésmodellezésének*, a kiejtett alak automatikus előállításának vizsgálata spontán nyelvű beszéd felismeréséhez. Azaz, hogy az esetlegesen többféle ejtésmódú vagy kivételes ejtésű szavak miként modellezhetők, a fonologikus átiratkészítés hogyan automatizálható.

Ezeken felül természetesen a nyelvi modellezés is fontos – látszólag erősen nyelvfüggő – feladat, azonban a megfelelő lexikai egységek kiválasztása után a standard N-gram modelleken túlmutató megközelítés kidolgozása nem látszott feltétlenül szükségesnek. Hasonlóan, a szerző a fizikai szintű akusztikai modellezést sem tartja (a tonális nyelvektől eltekintve) nyelvspecifikusnak, ezért annak részleteivel az értekezés nem foglalkozik.

[8] Gépi beszédfelismerés alatt az általános beszéd-szöveg átalakítást értjük, amely a nagyszótáras folyamatos spontán nyelvű beszéd szöveggé alakítását is magában foglalja.

4. A fonetikai koartikuláció modellezése magyar nyelvű beszéd gépi felismeréséhez

A fonetikai koartikuláció – az egymást követő beszédhangok egymásra hatása, „együtt ejtése” – a beszéd alapvető jellegzetessége. Nem különbözik e tekintetben a magyar nyelv más nyelvektől, amit az is jelez, hogy az elismert magyar nyelvű gépi *beszédelőállítási* módszerek mindegyike elemi vagy magasabb szinten explicit hangkapcsolati modelleket használ (diádok, triádok stb. [Olaszy & Németh+ 92], [Olaszy & Németh+ 00]).

A magyar nyelvű gépi beszédfelismerésnél ugyanakkor a fonetikai koartikulációs jelenségek explicit modellezésének szükségessége sokáig nem tűnt nyilvánvalónak. A kutatócsoportunkhoz kötődő publikációkat nem számítva, még a legutóbbi időkben is szinte kizárólagos a környezetfüggetlen modellek használatával elért eredmények publikálása [Tóth & Kocsor+ 04], [Vicsi & Velkei+ 05], [Tóth 06], [Bánhalmi & Paczolay+ 06], [Szaszák & Vicsi 07], [Tóth 09].

A következőkben két módszert vizsgálunk a magyar nyelvű fonetikai koartikuláció explicit, beszédhangonként három fázisban történő modellezésére. Az egyik esetben a beszédhangok környezetfüggő leképzése elemi (HMM állapot) szintre akusztikai információ felhasználása nélkül, alapvetően fonológiai szakértői szabályok alapján történik, melyet egyszerű statisztikai módszerrel korrigálunk (szabály alapú visszametszéses modell) [B2]. A másik esetben alapvetően akusztikai statisztikák alapján történik a leképezés kialakítása (ML döntési fák révén), csupán néhány (akár véletlenszerű) fonetikai kategóriát szükséges definiálni [Young & Odell+ 94]. Ez utóbbi megközelítést magyar nyelvre először [Szarvas & Furui 02]-ben alkalmazták. Referenciaként a környezetfüggetlen beszédhangmodellezést, azaz az implicit fonetikai koartikulációs modellezést is alkalmazzuk.

4.1. Bevezetés

A gépi beszédfelismerés hatásfokát alapvetően befolyásolja az alkalmazott beszédhangmodellek kialakításának módja. A legegyszerűbbek az ún. monofón modellek, amelyeknél fonémánként egyetlen beszédhangmodellt használunk. E megközelítés előnye nyilvánvaló, hiszen általában alig negyven beszédhangmodell elegendő például az európai nyelvek beszédakusztikájának teljes lefedésére (nyelvenként). A hátránya is e helyen keresendő: igen nehéz a beszéd óriási változatosságát ilyen kis számú beszédhangmodellel leírni.

Bár az egyes beszédhangmodellek komplexitásának növelése segíthet a beszédfelismerési pontosság növelésében, az általános tapasztalat mégis az, hogy nagyobb számú modell alkalmazása teszi alapvetően használhatóbbá, pontosabbá a beszédfelismerő rendszereket. A legelterjedtebb megközelítés a beszédhangok fonetikai – fonológiai környezetének figyelembevételével történő modellalkotás. Ezen belül szinte egyeduralkodó az a szemlélet, miszerint csak az adott beszédhangot közvetlenül megelőző és közvetlenül követő beszédhangot, mint fonetikai környezetet veszik figyelembe. Ezt ún. trifón modellezésnek szokták hívni. Például a „pamacs” szó első [a] hangját a „p-a+m” (általánosított) trifón modellezi, ahol a „p-” a bal kontextust, a „+m” a jobb kontextust jelöli. Fontos, hogy maguk a kontextusok fizikailag nem tartoznak a trifónhoz, csak a modellezett hang.

A trifón modellezés előnye, hogy mind a fonetikai mind a fonológiai koartikulációs jelenségek többségét – definíciójából eredően – jól képes kezelni [Jurafsky & Ward+ 01], és ez a felismerési pontosságban is tükröződhet. Hátránya viszont, szintén a definíciójából adódóan, hogy jóval bonyolultabb módon lehet általános felismerési hálózatokat építeni belőlük, mint monofón modellekből. További nehézséget okoz a modellek finomszerkezetének, paramétereinek meghatározása, vagyis az akusztikus modell tanítása.

A továbbiakban a trifón modellezés problémáival, lehetséges megoldásaival, valamint az ezekkel elért beszédfelismerési eredmények összehasonlító elemzésével foglalkozunk.

4.2. A trifón modellezés problémái

4.2.1. Tanítás

Problémákat vet fel, ha a környezetfüggő beszédhangmodelleknél minden fonológiai környezetet megkülönböztetünk, azaz *általánosított trifón modelleket* használunk. Ez könnyen szemléltethető a következő példával. Ha egy adott nyelv fonémáinak számát P=40-nek tekintjük, akkor a fonológiai szabályszerűségeket figyelmen kívül hagyva, elméletileg P^3, azaz 64.000 környezetfüggő beszédhangmodellt kapunk, hiszen a 40 beszédhang mindegyike elvileg 40x40-féle fonológiai környezetben valósulhat meg. Kimutatható, hogy még a hangsorépítési szabályok [Siptár 95] figyelembevételével is több ezer különböző általánosított trifón modell adódik. Elméletileg ugyan a modellek nagy száma kedvező a beszédváltozatosság lefedése szempontjából, a gyakorlatban azonban több problémát vet fel. Ugyanis minden beszédhangmodell betanításához legalább néhány száz mintára van szükség a modellezett hangból. Ez a fenti példában szereplő általánosított trifón modellek esetén azt jelenti, hogy még a legritkább általánosított trifón modellnél, mint pl. az „a-g+ny” esetén is több száz, lehetőleg természetes beszédben előforduló példát kellene használnunk a tanítási eljárás során. Sőt, ha beszélőfüggetlen felismerőt szeretnénk, arra sem árt ügyelni, hogy a beszélők nemben, korban változatosak legyenek.

Látható, hogy amennyiben minden fonológiai környezetet különbözőnek tekintünk, a trifón modellek betanításához olyan méretű és minőségű beszédadatbázisra lenne szükségünk, ami a világon szinte sehol nem áll rendelkezésre.

A megoldást az egyes (általánosított) trifón modellek, illetve állapotaik fonetikus környezetük alapján történő összevonása, illetve a modellállapotok számának az adott beszédadatbázishoz való illesztése jelenti. A trifónok, illetve állapotaik összevonására-csoportosítására számos megközelítés született. A legegyszerűbb, ha azokat az általánosított trifónokat (vagy trifón állapotokat), melyekre nem jutott elegendő tanítóminta monofón modellekkel helyettesítjük, ilyen megoldást alkalmaznak pl. a [Siivola & Hirsimäki+ 03]-ben. [Schramm 06] szerint jobb eredmények érhetők el szofisztikáltabb, pl. Gauss komponens összevonásokkal [Bellegarda & Nahamoo 90], vagy szabály alapú trifón csoportosítással [Lee & Giachin+ 90]. Kétségkívül a legkedveltebb megoldás a ML fonetikus döntési fákon alapuló trifón állapotcsoportosítás [Bahl & de Souza+ 93], [Young & Odell+ 94]. Ez utóbbit röviden ismertetjük, illetve összevetjük az általunk kidolgozott fonológiai szabály alapú visszametszéses trifón állapotcsoportosítással, melyet szintén a 4.3. alfejezetben mutatunk be.

4.2.2. Felismerési hálózatépítés

Monofón modelleknél a felismerési hálózat építése viszonylag egyszerű, hiszen minden szó helyébe elegendő a fonetikus átiratának megfelelő monofón modellek sorozatát illeszteni. Trifón modelleknél viszont – különösen a folyamatos beszédfelismerési hálózatok esetén – azokban a csomópontokban, ahol több szót több szó követhet, nem triviális feladat a szószintről a trifón hálózati szintre történő lépés. Ekkor ugyanis attól függően kell az egyik vagy másik trifón modellt használni egy adott szó végén, hogy az őt követő szó milyen fonémával kezdődik.

A probléma szemléltetése – a példában feltételezve, hogy a beszéd szünet után indul, az első szó után viszont nem tart szünetet a beszélő:

Pamacs a … →
sil-p+a p-a+m a-m+a m-a+cs a-cs+**a** cs-a+…

Pamacs egy …→
sil-p+a p-a+m a-m+a m-a+cs a-cs+**e** cs-e+gy e-gy+…

(„sil” jellel a beszédszünetet jelöltük, melyet fonetikus környezet értékűnek tekintünk.)

A problémának kétféle kezelése terjedt el.

1. A hálózatépítés egyszerűsítése és/vagy egyéb okokból a szószéli trifónok környezet-függésének redukciója. Ilyenkor, mivel nem tudjuk milyen a szókezdet esetén a bal-, szóvég esetén a jobboldali környezet, szó elején jobboldali difón modellt, szó végén pedig baloldali difón modellt használunk.

A szóbelsei trifón modellezés szemléltetése:

Pamacs a … →
p+a p-a+m a-m+a m-a+cs a-cs a …

Pamacs egy …→
p+a p-a+m a-m+a m-a+cs a-cs e+gy e-gy …

Ezt a megközelítést az angol nyelvű szakirodalom „word internal” trifón modellezésnek hívja, amit magyarul „*szóbelsejei”* trifón modellezésnek hívunk. Hátránya, hogy a szóhatárokon fellépő koartikulációs jelenségeket csak implicite, monofón szinten kezeli.

2. A korrekt, szóhatárokon is megfelelő modellek alkalmazását angolul „cross-word” trifón modellezésnek hívják, amit talán „*szóhatárokon átívelő”* trifón modellezésnek fordíthatunk. Ez a megközelítés – köszönhetően a szóhatárokon is explicit koartikulációs modellezésnek – az általános tapasztalatok szerint pontosabb beszédfelismerést tesz lehetővé, mint az előbbi módszer pl. [Aubert 99]. A felismerési hálózatépítés a 2. fejezetben említett véges átalakítók kompozíciójával hatékonyan megoldható [Mohri & Riley+ 98].

4.3. Trifón állapotcsoportosítási eljárások

Az „arany középút” a monofón és az általánosított trifón szemlélet között tehát a környezetfüggő beszédhangok, ill. állapotaik fonetikus környezetük alapján történő csoportosítása és összevonása egy „csoport” modellé. A csoportosítást célszerű trifón állapotonként végezni, így a kiejtést három fázisra bontva magasabb szinten optimializált modelleket kaphatunk, mintha beszédhang szinten tennénk ugyanezt.[9] Ezeket a beszédhangmodelleket nevezzük állapotcsoportosított („state clustered”) trifón modellnek. A továbbiakban az „állapotcsoportosított” jelzőt általában elhagyjuk.

A feladat tehát a környezetfüggő beszédhangmodellek oly módon való kialakítása, hogy azok egyrészt minél jobban fedjék le az adatbázis fonetikai–fonológiai gazdagságát, másrészt, hogy elegendő mennyiségű mintával legyen tanítva minden egyes beszédhangmodell.

A megoldás nem triviális, hiszen a két követelmény, hogy legyen minél több modellünk, ugyanakkor az egyes modelleket minél több mintával tanítsunk, adott adatbázis esetén ellentmond egymásnak. Az ellentét úgy oldható fel, hogy a modell komplexitást a rendelkezésre álló adatbázis méretéhez igazítjuk.

A következőkben két módszert ismertetünk trifón beszédhangmodellek állapotainak csoportosítására.

4.3.1. Visszametszéses fonológiai trifón állapotcsoportosítás

A módszer azért kapta „visszametszéses” jelzőt, mert a környezetfüggőség mértékét a nyelvi modelleknél használt visszametszéses („back-off”) megközelítéshez hasonlóan redukálja. A „fonológiai” jelzőt pedig azért, mert a környezetfüggő beszédhangmodell állapotok csoportosítására nem használ fel más információt, mint az adatbázis fonológiai statisztikáit és szabályokat. A gyakorlatban ez azt (is) jelenti, hogy a tanító-beszédadatbázis két alapvető része, vagyis a hullámformák és a szöveges átiratok közül csak az utóbbi alapján történik a csoportok kialakítása. A kísérletekben az egyszerűség kedvéért BO-trifónokként (BO: Back-Off) fogunk a következőkben részletezett módon kialakított környezetfüggő beszédhangmodellekre.

A csoportok kialakításának a módja a következő:

1. Az első lépés a bal ill. jobbkörnyezetek definiálása és azok hierarchikus sorba rendezése. Ezeket a továbbiakban hierarchikus szabályoknak nevezzük, melyek a bemenő általánosított trifón állapotok elsődleges csoportokba foglalását végzik. A környezetdefiníciók kialakítása és a sorba rendezés szakértői feladat.

   Szemléltetésül legyen a bemenő általánosított trifón állapot:

   p-a+m_s1: a „p” bal- és „m” jobbkörnyezetű „a” beszédhangmodell első, azaz bal szélső állapota

[9] Mind a trifón, mind a monofón beszédhang modelleket 3 állapotú HMM-mel reprezentáljuk.

Példa a hierarchikus bal és jobb környezetekkel megadott szabályokra:

NAZÁLIS: m, n, …
ALVEOLÁRIS: d, t, n, …
VELÁRIS: g, k, …
BILABIÁLIS: p, b, m…
EGYÉB: *

A fenti szabályrendszert alkalmazva a “pamacs” szó első [a] hangjának s1 állapotára (p-a+m_s1) a „BILABIÁLIS-a+NAZÁLIS_s1” trifón állapotot kapjuk. Természetesen egy fonéma – mint környezet – több szabálynál is szerepelhet, a bemenő adatra csak az vonatkozik, amelyik magasabb hierarchiaszinten van. (Az utolsó sorban szereplő EGYÉB szabály minden hangkörnyezetre illeszkedik. A szabályok előtti Bal_ ill. Jobb_ tag megadása nem kötelező, ilyenkor mindkét oldali környezetre érvényes a szabály.)

2. Az összes elvi általánosított trifónra alkalmazzuk az 1. szabályrendszert, így kialakulnak az elvi csoport trifón állapotok.

   Illusztráció: BILABIÁLIS-a+NAZÁLIS_s1

3. Feltételezzük, hogy rendelkezésre áll az elemi akusztikus modellek tanításához használt beszédadatbázis a hozzá tartozó fonémasorozattal. Ez alapján az általánosított trifón sorozat triviálisan előállítható. Ezután az egyes általánosított trifónokat állapotokra képezzük le a fenti szabályrendszer segítségével. Végül statisztikát készítünk arról, hogy a 2. pontban előállított egyes elvi csoport trifón állapotokra mennyi tanítóminta (fonémaszámban) adódik.

   Illusztráció: BILABIÁLIS-a+NAZÁLIS_s1 : 7 db

4. A szélső és középső állapotokra külön definiált vágási küszöb alatti tanítómintaszámú elvi trifón állapotokat az 1.5.1. -ben ismertetett módszerhez hasonlóan visszametsszük difón-szerű állapottá (kivéve a már difón vagy monofón-szerű állaptokat). Difón-szerű állapotnak az egyik oldalán EGYÉB csoportosítású állapotot, monofón-szerűnek a mindkét oldalán ilyen jelzésű állapotot nevezzük.

   Illusztráció: BILABIÁLIS-a+NAZÁLIS_s1 : 7 db < 10 db // szélső trifón küszöb
   EGYÉB-a+NAZÁLIS_s1 : 31 db
   BILABIÁLIS-a+EGYÉB_s1 : 5 db →

   EGYÉB-a+NAZÁLIS_s1 : 31 db
   BILABIÁLIS-a+EGYÉB_s1 : 12 db

   Trifón állapotoknál azt a visszametszési szabályt alkalmazzuk, hogy a szélső állapotoknál mindig az ellenoldali környezetet vágjuk le, a középső állapotnál pedig arra az oldali difónra-szerű állapotra metsszük vissza a trifón állapotot, amelyik előfordulása eleve nagyobb volt.

5. Azokat a difón-szerű állapotokat, melyek az adott difón vágási küszöb alatti előfordulással bírnak, visszametsszük monofón-szerű állapottá.

A fenti módszer nem biztosítja minden szabályrendszer esetén, hogy a monofón-szerű trifón állapotokra jut elég tanítóminta. Ezért elvileg szükséges lehet valódi monofón állapotokkal történő helyettesítésük. Tapasztalataink szerint azonban a gyakorlatban előforduló szabályrendszerek esetén erre általában nincs szükség.

Az eljárás előnye, hogy könnyen implementálható, nagyon gyors és nem szükséges hozzá semmilyen előzetes akusztikai modelltanítás, mivel csak a tanító-adatbázis fonéma szintű információira épít. Hátránya, hogy a nagy szakértelmet igénylő szabályhierarchia kialakítás a kutató-fejlesztő feladata. A végeredményül előálló össz. állapotszám a trifón és difón vágási küszöbökkel skálázható bizonyos határok között.

4.3.2. Fonetikus ML döntési fa alapú trifón állapotcsoportosítás

Ez a széles körben használt megközelítés [Young & Odell+ 94] akusztikus-fonetikai információt is használ a trifón állapotcsoportok kialakításához. Bemenetként az akusztikai előfeldolgozáson átesett tanító-adatbázist, a hozzá tartozó fonémasorozatot, illetve fonetikai kategóriákat vár, a leképezési szabályokat ezekből ML (Maximum Likelihood) döntési fák építésével automatikusan alakítja ki.

A fonetikai kategóriák megadása tetszőleges sorrendben történhet, illetve tetszőleges új szabály nem ronthatja a tanító-adatbázishoz való illeszkedés mértékét. Bár általában egyszerűbb e kategóriákat alapszintű fonetikai ismeretek alapján „kézzel” megadni, az előző tulajdonságok jól algoritmizálhatóvá teszik magát a fonetikai kategóriakialakítást is. Például [Beulen & Ney 98] és [Singh & Raj+ 99] is kézi, szakértői kategóriadefiníciónál jobb (illetve nem rosszabb), automatikus módszert mutat be különböző nyelvek fonetikai kategóriáinak meghatározására.

Egy adott beszédhang trifón adott állapotának csoportjait lépésről lépésre határozzuk meg. Kezdetben az összes – adott hangrészlethez tartozó – trifón állapotot egy csoportba tartozónak tekintjük, majd egy fonetikai kategóriát a jobb- vagy baloldali környezetre való kérdésként értelmezve, két csoportra osztjuk a trifón állapotokat. A módszer lényege, hogy lépésenként azzal a szabállyal osztjuk ketté az arra legalkalmasabb csoportot, amelyiknek az alkalmazásával az állapotcsoportosított trifón modellek tanító-adatbázishoz való illeszkedésének mértékét leginkább növeljük. Végeredményben így minden fonéma beszédhangmodelljének mindhárom állapotához egy fonetikus döntési fa áll elő, amelynek levelei reprezentálják az adott hangrészlethez tartozó trifón állapotcsoportokat. A későbbiekben a DT-trifón (DT: Decision Tree) rövid névvel fogunk hivatkozni ezekre a beszédhangmodellekre.

A trifón állapotcsoportok kialakításának menete a következő:

1. Az első lépés a bal és jobbkörnyezetek, mint fonetikai kategóriák definiálása. Az előző eljárással ellentétben ezeket nem kell sorba rendezni, és a továbbiakban fonetikus környezetre utaló kérdésekként fogunk rájuk hivatkozni.

   NAZÁLIS: m, n, …
   ALVEOLÁRIS: d, t, n, …
   VELÁRIS: g, k, …
   BILABIÁLIS: p, b, m…

2. Az előző lépéssel párhuzamosan történhet az akusztikai szintű tanítás, statisztika-készítés. Ez azt jelenti, hogy a tanító-adatbázishoz tartozó jellemzővektor- és fonémasorozat alapján el kell készíteni az általánosított trifón szintet, és azok 3 állapotú rejtett Markov modelljeit állapotonkét 1 Gauss komponenssel be kell tanítani. A tanítás során az egyes elemi akusztikus modellek előfordulási számait és a tanító-adatbázison mért hasonlósági mértékeiket rögzíteni kell.

3. A fonetikus kérdések és a statisztikák alapján ML döntési fával alakítjuk ki a fonéma és állapot típusonként a trifón állapotcsoportokat a következő módon:

   a. Inicializálás: úgy tekintjük, hogy az adott fonéma adott állapotát, pl. az „a” bal szélső állapotát egyetlen csoporttal modellezzük melybe az összes, az adott hanghoz tartozó általánosított trifón modellállapot belekerül. Ez a döntési fa kezdőpontja. A statisztikákból meghatározható, hogy az ebbe a csoportba foglalt állapotok mennyivel járulnak hozzá eloszlásaik révén az együttes tanító-adatbázison mért hasonlósági mértékhez.

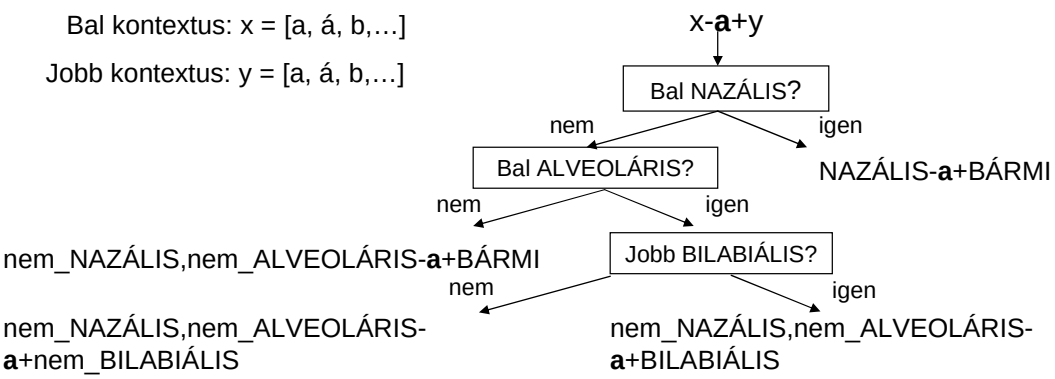
   b. Kérdések feltevése: kezdetben a kezdőpontban, később minden végpontban feltesszük az összes kérdést a bal ill. jobbkörnyezetre vonatkozóan. Minden kérdés két csoportra bontja azt a csoportot, amelyikre feltettük: az egyikbe azok az általánosított trifón állapotok kerülnek melyeknél igen a válasz, a másikba a többi.

   c. Kérdések kiértékelése: bármelyik végpontot bármelyik kérdés alapján két csoportra bontunk, az növelheti az együttes hasonlósági mértéket. Ez a növekmény a 2. pontbeli statisztikák alapján minden kérdésfeltevésnél kiszámítható.

   d. Fa növelése vagy a növelés elvetése: amelyik végpontban és amelyik kérdésnél a hasonlósági mérték növekedése a legnagyobb volt, azt választjuk a fa növelésére. Ha azonban a régi helyett így keletezett két új csoport egyikére jutó tanítóminták száma egy küszöb alá esne, nem hajtjuk végre a döntési fa-növelést. Hasonlóan, definiálható egy hasonlósági mérték növekedési küszöb is, amely alatti növekedés esetén szintén nem növeljük tovább a fát.

Tehát a b, c, d lépések ciklikus ismétlésével kialakul egy döntési fa, melynek végpontjain – „levelein” – vannak a kívánt trifón állapotcsoportok.

Illusztráció – egy lehetséges döntési fa az „a” hang s1 állapotára:



4.1. ábra. A fonetikai döntési fa alapú trifón állapotcsoportosítás szemléltetése.

Mivel a bal, középső és jobb állapotokra külön történhet a döntési fa építése, ezért a hasonlósági mérték növekedést a legalacsonyabb szinten optimalizáltuk.

Az eljárás szépsége, és gyakorlati szempontból óriási előnye, hogy csak a kérdések definiálása a fejlesztő feladata (ami akár nyelvészeti ismeretek nélkül, statisztikai módszerekkel is történhet), a teljes, és az előző módszernél akár jóval összetettebb szabályhierarchia automatikusan direkt ML optimalizálás során alakul ki.

További előnye az eljárásnak, hogy a használójának csak két küszöböt kell megadnia, azokkal tudja vezérelni a fa leveleinek számát, illetve, hogy új, tetszőleges szabály hozzáadása nem ronthatja a tanító-adatbázishoz való illeszkedés jóságát.

A módszer hátránya, hogy el kell végezni hozzá az általánosított trifón modellek betanítását, és hogy a széles körben alkalmazott [Young 06] eszköz csak 1 Gauss modell / állapottal jellemzett modellek esetén képes elvégezni a döntési fa építést.

Szemléltetésül, például a „pamacs” szó első „a” hangjának a (csoport) trifón modellje a szabályok és a döntési fa alapján a „nem_NAZÁLIS,nem_ALVEOLÁRIS-**a**+BÁRMI” lesz, míg a második „a” hangja a „NAZÁLIS-**a**+BÁRMI” trifón csoportba kerül.

4.4. Gépi beszédfelismerési kísérletek

A statisztikai módszereken alapuló beszédfelismerési kísérleteinkben azt vizsgáltuk, hogy különféle tanítási és tesztelési konfigurációkban hogyan változik a felismerés pontossága. Minden konfigurációban három akusztikus modellezési megközelítést vizsgáltunk: a monofón modellezést, a visszametszéses fonológiai állapotcsoportosítású és a fonetikus a döntési fa alapú trifón modellezést.

4.4.1. Beszédadatbázisok

A tanító és tesztelő adatbázisokat a legnagyobb, nagyrészt publikus magyar telefonos beszédadatbázisok, az MTBA [Vicsi & Tóth 02], a Besztel, a SpeechDat és a Tesztel [Vicsi et al.] összességéből alakítottuk ki. Ezek az adatbázisok elsősorban olvasott beszédet, valamint kisebb arányban spontán bemondásokat is tartalmaznak. Az adatbázisok jól tervezettek, mind korban, nemben és hangkapcsolat-statisztikákban jól reprezentálják a magyar nyelvű beszédet. Mindegyik adatbázis tartalmaz fonetikailag változatos szavakat és mondatokat, valamint a parancsszavas vezérlés során gyakran előforduló „izoláltan" ejtett kifejezéseket. Az első három adatbázis lényegében ugyanarra a szövegkorpuszra épül, és mindegyiknek az általunk elérhető része 500 beszélőtől tartalmaz hanganyagot. A Tesztel adatbázis 100 beszélős, és jellegzetessége, hogy szándékosan nagy és természetes háttérzajban felvett bemondásokat tartalmaz. Az adatbázisokban a vonalas és mobil telefonos felvételek összességében körülbelül ugyanolyan számban képviseltetik magukat.

Tanítóhalmazok: Tanítás céljára az MTBA, Besztel, és a SpeechDat adatbázis 500-400-400 beszélőjének azon felvételeit jelöltük ki, melyek nem „o", és „z" jelzésűek, azaz nem tartalmaznak tulajdonneveket és bizonyos típusú mondatokat. A SpeechDat esetén csak egy szűkebb halmazt, a fonetikailag változatos mondatokat és szavakat használtuk.

A teljes tanítóhalmaz mellett annak bizonyos részhalmazait is képeztük, hogy az egyes akusztikus modellezési eljárások tanító-adatbázismérettől való függését is vizsgálhassuk.

A tanítóhalmazok jelölése és tartalma:

- **M**: Az MTBA fonetikailag változatos mondatai és szavai, kivéve a „z" jelzésű mondatok, 500 beszélő, 6000 felvétel
- **MM**: Az MTBA összes tanító felvétele, 500 beszélő, 19000 felvétel.
- **MM_BS**: Az MTBA és a Besztel összes tanító felvétele, 900 beszélő, 39000 felvétel.
- **MM_BS_SD**: Az MTBA, a Besztel és a SpeechDat tanítófelvételei, 1300 beszélő, 44000 felvétel.

Teszthalmazok:A teszthalmazokat úgy állítottuk össze, ne legyen átfedés a tanítóhalmazban szereplő beszélőkkel. Így a tanításnál fel nem használt felvételek összesen 220 beszélőtől (Besztel 100, Speechdat 100, Tesztel 20) kerültek a teszthalmazokba.

Alapvetően kétféle felismerési tesztet végeztünk a tesztelő felvételek típusától függően. Az izolált szavakat, kifejezéseket tartalmazó bemondásokat (2475 felvétel) izolált szavas felismerővel ismertettük fel, a mondatokat (2385 felvétel) pedig folyamatos beszédfelismerővel. Számokkal, dátumokkal és egyéb, kapcsolt szavas felismeréshez illeszkedő felvételekkel nem teszteltünk.

Az *izolált szavas* halmaz azon felvételeit, amelyek lexikálisan illeszkedtek a tanítóhalmazhoz, azaz a tesztelő szóalakok szövegszerűen szerepeltek a tanítás során is (pl. fonetikalilag

változatos szavak), az „illeszkedő” teszthalmazba tettük. Értelemszerűen a többi felvételt (mint például a tulajdonnevek) a „nem illeszkedő” teszthalmazba tettük.

A *folyamatos* beszédfelismerési teszteknél nem csak lexikális, hanem nyelvi illeszkedésről is beszélhetünk. Az egyik halmazba azokat a mondatokat válogattuk, amelyeknek szövegét mind az akusztikus, mind a nyelvi modell tanításakor felhasználtunk, ez az „illeszkedő” halmaz (lexikális és nyelvtani szempontból is). A másik, „nem illeszkedő” teszthalmazba azok a felvételek kerültek, melyeknek szövege sem az akusztikus, sem a nyelvi modell tanításakor sem lett felhasználva. Egyéb halmazt nem vizsgáltunk.

A teszthalmazok jelölése és tartalma:

- **I_M**: Izolált szavakat, kifejezéseket tartalmazó, a tanítóadatokhoz lexikálisan illeszkedő felvételek, 220 beszélő, 1726 felvétel.
- **I_U**: Izolált szavakat, kifejezéseket tartalmazó, a tanítóadatokhoz lexikálisan nem illeszkedő felvételek, 220 beszélő, 749 felvétel.
- **C_M**: Nyelvi és lexikális szempontból a tanításhoz illeszkedő mondatok, 220 beszélő, 1973 felvétel („s” jelzésű mondatok a Besztel-ből és a SpeechDat-ból, „s1” és „s2” jelzésű mondatok a TeszTel-ből).
- **C_U**: Sem nyelvi és sem fonológiai szempontból a tanításhoz nem illeszkedő mondatok, 220 beszélő, 412 felvétel („z” jelzésű mondatok a Besztel-ből és a SpeechDat-ból, „s3” jelzésű mondatok a TeszTel-ből).

4.4.2. Beszédfelismerési paraméterek, beállítások

Lényegkiemelés: Lényegkiemelési paraméterekként a bemenő beszédjelből MFCC (Mel Frequency Cepstral Coefficients) 12 dimenziós vektorokat képeztünk, melyekhez *logE* (keretenkénti logaritmikus energia) paramétert is csatoltunk, majd dinamikus Delta és Delta-Delta értékeket számítottunk. A statikus energiát végül kicsatolva összesen 38 dimenziós jellemzővektorokat kaptunk. Mind a tanítás, mind a tesztelés során alkalmaztuk a vak csatornakiegyenlítés (Blind Equalization) módszerét. A kísérletekben a [C14]-ben bemutatott lényegkiemelő eszközt alkalmaztuk.

Elemi akusztikus modellek: Az atomi modellek rejtett Markov-modell állapotok voltak, rögzített hurok és továbblépési valószínűségekkel. Állapotonként maximum 10 Gauss függvényből álló folyamatos megfigyelési sűrűségfüggvényeket használtunk. A tanítás részben a [J6]-ban bemutatott, részben a HTK (Hidden Markov-Model Toolkit) eszközzel [Young 06] történt.

Fonetikai koartikulációs modellek: Mind a monofón mind a trifón modelleknél a beszédhangokat 3 elemi akusztikus modellre képeztük le. Az előbbi esetben a környezettől függetlenül, az utóbbi esetben a visszametszéses módszer [B2], ill. az ML döntési fa alapján a fonetikus környezettől függően [Young 06]. A trifón állapotcsoportosításokat tanítóhalmazonként külön-külön végeztük.

Környezetfüggőségi modell*:* szóhatárokon átívelő trifón modellezés.

Fonológiai koartikulációs modell: *explicit modellt* nem alkalmaztunk, azaz implicite, fonetikai szinten volt kezelve.

Szótármodell: A kiejtési modellek nyers, fonológiai koartikulációkat nem tartalmazó fonemikus átiratait automatikusan, graféma-fonéma szabályok segítségével állítottuk elő [J2, J6, B3, B4]. Allofónikus változatokat nem jelöltünk, továbbá a hosszú és rövid mássalhangzókat sem különböztettük meg. Így – a szünetmodelleket nem számítva – összesen 39 fonológiai kategóriát használtunk. A szünetmodell háromállapotú környezetfüggetlen modell volt.

Az izolált szavas felismeréseknél ugyanazt az 1334 elemű szótárat használtuk az illeszkedő és nem illeszkedő felvételek esetén is. Hasonlóan, a folyamatos felismeréseknél is ugyanazt az 5561 elemű szótárat és természetesen ugyanazt a nyelvi modellt alkalmaztuk mindkét teszthalmaz esetén. Mind a folyamatos mind az izolált szavas tesztek esetén a teljes teszthalmazt lefedő szótárakat alkalmaztuk, így szótáron kívüli elemek kezelésére nem volt szükség.

Nyelvi modell: A folyamatos felismerésnél szó-trigram nyelvi modelleket alkalmaztunk Katz-féle visszametszéssel [Katz 87] és Good-Turing valószínűség-újraelosztással [Good 53]. A tanítószöveg az illeszkedő tesztmondatok szövege alapján készült úgy, hogy minden különböző mondatot csak *egyszer* szerepeltettünk. Így az illeszkedő mondatokon PP=40-es perplexitást [Bahl & Jelinek+ 83], a nem illeszkedő tesztmondatokon PP=6230-as (nagyon magas, azaz igen kedvezőtlen) perplexitás értéket kaptunk. A nyelvi modellezésre az SRILM (Stanford Research Institute Language Modeling) eszközt alkalmaztuk [Stolcke 02].

A beszédfelismerési tudásforrások integrációját és optimalizációját a 2. fejezetben ismertetett WFST módszerekkel végeztük, ehhez a AT&T FSM Toolkit-jét használtuk[10]. A felismerő motor a [C7]-ban említett, optimalizált Viterbi-algoritmuson (dinamikus programozáson) alapuló eszköz volt. A keresési mélységet fixen úgy állítottuk be, hogy a felismerési pontosság minden esetben bőven a telítési szakaszra essen (azaz, az a praktikusnál jóval nagyobb keresési teret vizsgáltunk), így különböző RTF mellett is összehasonlíthatók az eredmények.

4.4.3. Az akusztikus modelltanítás eredményei

Az akusztikus modell tanításának eredménye alatt adott tanító-adatbázis mellett létrejött (elemi) beszédhangmodellek összességét értjük. Ezek komplexitására utal az összesített állapotszám, mely közvetlenül befolyásolja a rendszer tárkapacitás igényét és közvetetten a felismerési folyamat sebességét. A két trifón akusztikus modelltípus kialakításánál arra törekedtünk, hogy a felismerési pontosság maximális legyen.[11]

4.1. táblázat. A monofón (Monofón), a visszametszéses fonológiai csoportosítású trifón (BO-Trifón) és a fonetikai döntési fa csoportosítású trifón (DT-Trifón) akusztikai modellek állapotszáma a tanítóhalmaz függvényében.

Állapotszám:	*M*	*MM*	*MM_BS*	*MM_BS_SD*
Monofón	126	126	126	126
BO-Trifón	998	1342	1706	1904
DT-Trifón	1775	2732	3952	4618

[10] http://www2.research.att.com/~fsmtools/fsm/

[11] Végeztünk ellenőrző kísérleteket azonos trifón állapotszámra törekedve a két konkurens állapotcsoportosítás között, ezek eredményei azonban lényegükben nem különböztek a következőkben bemutatandóktól.

4.4.4. Izolált szavas beszédfelismerési eredmények

Az izolált szavas szófelismerési arányok a 4.2. és 4.3. táblázatban találhatók. Minden esetben az előbb részletezett, összesen 12 akusztikus modellt használva születtek az eredmények.

4.2. táblázat. Izolált szavas szófelismerési [%] arányok *illeszkedő* teszthalmaz esetén.

I_M Corr:	*M*	*MM*	*MM_BS*	*MM_BS_SD*
Monofón	82,91	86,33	86,73	86,73
BO-Trifón	*92,99*	*94,9*	*96,06*	*95,94*
DT-Trifón	*94,79*	*96,87*	*96,87*	*96,58*

4.3. táblázat. Izolált szavas szófelismerési [%] arányok *nem illeszkedő* teszthalmaz esetén.

I_U Corr:	*M*	*MM*	*MM_BS*	*MM_BS_SD*
Monofón	80,61	82,49	84,09	83,42
BO-Trifón	*90,91*	*91,41*	*91,44*	*92,11*
DT-Trifón	*93,32*	*93,58*	*93,32*	*93,98*

Látható, hogy minden oszlopban a fonetikai csoportosítás eredménye a legjobb, attól kevéssel elmarad a fonológiai csoportosítású trifón modellé és két-háromszor akkora felismerési hibával követi a monofón eredmény. Az is megfigyelhető, hogy a tanító-adatbázis növekedésével általában javulnak az eredmények, de a várakozásoknál kisebb mértékben.

A várakozásoknak megfelelően pár százalékkal elmaradnak a tanítás során nem látott hangkapcsolatokat is tartalmazó teszthalmazon mért eredmények a másik (illeszkedő) halmaz eredményeitől. Érdekes, hogy minden konfigurációban szinte állandó 3% körüli eltérés figyelhető meg a két halmaz eredményei között. Az eltérések a három megközelítés eredményei között minden esetben szignifikánsak.

4.4. táblázat. Az izolált szavas gépi beszédfelismerés átlagos számításigénye (Real-Time Factor).

RTF:	*M*	*MM*	*MM_BS*	*MM_BS_SD*
Monofón	0,066	0,067	0,067	0,067
BO-Trifón	0,089	0,105	0,11	0,12
DT-Trifón	0,077	0,095	0,127	0,13

Az egyes akusztikai modellek gyakorlati alkalmazhatósága érdekében fontos a számításigényeiket is megvizsgálni. A 4.4. táblázat alapján azt mondhatjuk, hogy nincs nagyságrendi különbség az egyes megközelítések között e tekintetben.

4.4.5. Folyamatos beszédfelismerési eredmények

A folyamatos beszédfelismerésnél lényeges újdonság, hogy nyelvi modell alkalmazására is szükség van, melynek minősége nagymértékben befolyásolja a felismerési eredményt. A nyelvi modell használatának célja, hogy csökkentse az akusztikus modellre háruló döntés nehézségét azáltal, hogy valószínűségi becslést ad a felismerési szósorozatra – pusztán szövegstatisztikai alapon. A nyelvi modell PP értéke [Bahl & Jelinek+ 83] szemléletesen azt mutatja meg, hogy egy szó után átlagosan hány egyformán legvalószínűbb szó következhet.

Az 4.5. és 4.6. táblázatban láthatók a folyamatos felismerési eredmények. A jóval összetettebb feladat ellenére az illeszkedő tesztfelvételek esetén hasonló felismerési arányokat kaptunk, mint az izolált szavas teszteknél.

4.5. a) táblázat. A folyamatos beszédfelismerési tesztek szófelismerési [%] arányai *illeszkedő* teszthalmaz esetén.

C_M Corr:	*M*	*MM*	*MM_BS*	*MM_BS_SD*
Monofón	85,34	80,19	79,87	80,74
BO-Trifón	*92,72*	*91,43*	*92,15*	*92,84*
DT-Trifón	*94,13*	*93,82*	*94,22*	*94,47*

4.5. b) táblázat. A folyamatos beszédfelismerési tesztek szófelismerési [%] pontosságai *illeszkedő* teszthalmaz esetén.

C_M Acc:	*M*	*MM*	*MM_BS*	*MM_BS_SD*
Monofón	84,32	78,62	78,54	79,52
BO-Trifón	*91,37*	*89,84*	*90,58*	*91,27*
DT-Trifón	*92,54*	*91,97*	*92,55*	*92,93*

4.6. a) táblázat. A folyamatos beszédfelismerési tesztek szófelismerési [%] arányai *nem illeszkedő* teszthalmaz esetén.

C_U Corr:	*M*	*MM*	*MM_BS*	*MM_BS_SD*
Monofón	29,22	25,4	24,72	25,58
BO-Trifón	*52,16*	*50,3*	*53,63*	*54,88*
DT-Trifón	*61,95*	*61,27*	*64,24*	*64,42*

4.6. b) táblázat. A folyamatos beszédfelismerési tesztek szófelismerési % pontosságai *nem illeszkedő* teszthalmaz esetén.

C_U Acc:	*M*	*MM*	*MM_BS*	*MM_BS_SD*
Monofón	23,65	19,36	18,54	20,54
BO-Trifón	*41,66*	*39,37*	*42,05*	*43,48*
DT-Trifón	*48,16*	*47,66*	*52,34*	*52,02*

Drasztikus teljesítményromlást láthatunk viszont a nem illeszkedő felvételek esetén. Itt mutatkozik meg a nyelvi modellhez való illeszkedés jelentősége. A nagy különbség mindazonáltal jól indokolható a két teszthalmazon mért gyökeresen eltérő perplexitás (PP) értékekkel. Miként az izolált szavas teszteknél, itt is minden esetben szignifikáns a különbség a háromféle felismerési megközelítés eredményei között.

Fejezetünk talán legfontosabb megfigyelése a 4.6. táblázatból olvasható ki. Jól látható, hogy a nem illeszkedő teszthalmazon a monofón modellek folyamatos felismerési eredménye, azaz a felismerési általánosító képessége összehasonlíthatatlanul gyengébb a trifón modellekénél. Míg az egyéb teszteken a monofón felismerési hiba volt fele-harmada a trifón hibának, addig itt a felismerési *arány* feleződik-harmadolódik meg a trifón esethez képest. Vagyis a monofón modellek felismerési pontossága a korábbi hibaarányainak szintjére süllyedt (kb. minden 4.-5.

szó helyes csak!), míg a jobban teljesítő fonetikai csoportosítású trifón modellnél a szavak csaknem kétharmadát továbbra is helyesen ismertük fel.

Nem várt tapasztalat volt ugyanakkor, hogy a tanító-adatbázis méretének növelése alig javított a felismerési eredményeken. Itt a nagyobb tanítóhalmazoknál a tesztfelvételekhez képesti nagyobb fonológiai illesztetlenség, illetve az adatbázisok gyakorlatilag közös szövegkorpuszra épülése lehetnek a mögöttes okok.

4.7. táblázat. Az folyamatos gépi beszédfelismerés átlagos számításigénye (Real-Time Factor).

RTF:	*M*	*MM*	*MM_BS*	*MM_BS_SD*
Monofón	0,54	0,61	0,61	0,61
BO-Trifón	0,77	0,84	0,89	0,90
DT-Trifón	0,57	0,69	0,78	0,85

Ahogy a 4.7. táblázat mutatja, hogy a trifón modellek sokkal jobb folyamatos felismerési pontossága nem jár együtt hasonló mértékű számításiigény-növekedéssel. Sőt, még kisebbek a különbségek, mint az izolált szavas felismerésnél.

4.5. Összefoglalás

Bemutattunk egy új módszert környezetfüggő beszédhangmodellek kialakítására. Az eljárást visszametszéses fonológiai trifón állapotcsoportosításnak neveztük, mert pusztán fonológiai szintű szabályrendszer és statisztikák alapján történik a környezetfüggő beszédhangmodellek kialakítása akusztikai információ felhasználása nélkül.

A módszert összevetettük a környezetfüggetlen, illetve a fonetikai döntési fa alapú trifón állapotcsoportosítási eljárással. Számos beszédfelismerési kísérletet végeztünk nagy mennyiségű és több mint ezer beszélőtől származó tanító és teszt felvétellel. Izolált szavas és folyamatos beszédfelismerési teszteket egyaránt végeztünk különféle módon illeszkedő teszthalmazokkal.

Tapasztalataink szerint az inkább statisztika (ML döntési fák) alapján készült trifón modellek következetesen és szignifikánsan jobban teljesítenek, mint az inkább fonológiai szabályok alapján állapotcsoportosított környezetfüggő beszédhangmodellek. Ugyanakkor, a monofón modellezés eredményeihez képest drasztikusan jobb felismerési eredmények adódtak mindkét környezetfüggő beszédhangmodellezésen alapuló módszerrel.

Végkövetkeztetésünk, hogy mivel a környezetfüggő beszédhangmodellek minden vizsgált körülmény között lényegesen jobb felismerési hatásfokot biztosítanak mint a környezetfüggetlenek; használatuk magyar nyelv esetén is – különösen folyamatos beszédfelismerésnél – általánosító képességük miatt, úgymond „kötelező”.

5. A fonológiai koartikuláció modellezése magyar nyelvű beszéd gépi felismeréséhez

A fonológiai koartikulációk – magyarul főként kiejtési szabályokként is közismert fonológiai jelenségek – explicit modellezése szinte a beszédfelismerés kutatásának kezdetétől foglalkoztatta a kutatókat. Számos cikk, disszertáció foglalkozott a témával, azonban jelentős előrelépést nem sikerült elérni. Sőt, az utóbbi időkben mindinkább az explicit szabályoktól mentes, statisztikai alapú implicit kiejtési (és fonológiai koartikuláció) modellezési technikák nyernek tért.

Ebben a fejezetben egy rövid irodalmi áttekintéssel kezdünk, melynek célja szembesíteni az olvasót az ambivalens kutatási eredményekkel. Majd bemutatjuk a témával kapcsolatos munkánkat, mellyel reményeink szerint egy eddig hiányzó láncszemet pótlunk a „hagyományos" és az új, implicit kiejtésmodellezés között.

5.1. Bevezetés

Korábban a fonológiai koartikulációs jelenségek gépi beszédfelismerésnél történő (explicit) modellezésének nagy jelenséget tulajdonított a nemzetközi beszédkutató közösség. [Cohen 89] mind a magán-, mind a mássalhangzók ejtésvariációit alternatív allofón realizációkkal javasolta modellezni. Részben ennek nyomán igen elterjedt a fonológiai szintű alternatív kiejtési változatok alkalmazása, melynél a fonológiai koartikulációs és az egyéb (pl. nyelvjárási) eredetű kiejtési variációkat általában nem választották szét. Azonban mint [Lamel & Adda 96] rámutat, a túl sok alternatíva konfúzzá teszi a felismerési hálózatot, így a felismerési pontosság romlani fog. Különösen a szóhatárokon fellépő fonológiai koartikuláció modellezése problematikus a lexikai szinten, amire a megoldást a fonológiai koartikulációs szabályok véges állapotú gépekkel történő *környezetfüggő* modellezése és integrálása jelentette [Kaplan & Kay 94], [Mohri & Sproat 96]. [Hazen & Hetherington+ 02] súlyozott FST alapú fonológiai szabályreprezentáció mellett felismerési pontosság javulásról (4 – 8 %) számol be angol nyelvű telefonbeszéd-felismerés esetén. Nem világos azonban, hogy milyen részben nyelvjárási és milyen részben koartikulációs eredetű jelenségeket modelleztek a kísérletekben.

Ugyanakkor [Jurafsky & Ward+ 01] meggyőző kísérletekkel támasztja alá, hogy a szótagszintű kiejtésbeli megváltozásoknál kisebbeket – a fonológiai koartikulációk döntően ilyenek – a trifón modellezés (implicite) jól kezeli. Majd a nemzetközi kutatási trendek mind inkább az implicit kiejtésmodellezés felé irányulnak [Hain 02], [Kanthak & Ney 02], [Killer & Stüker+ 03], ami megkérdőjelezi az explicit fonológiai koartikulációkezelés szükségességét a statisztikai alapú gépi beszédfelismerésben. (A nem koartikulációból eredő fonológiai kiejtési variációk – pl. nyelvjárási, szleng, beszédhiba stb. eredetű változások – modellezését itt nem tárgyaljuk.)

A magyar nyelvi fonológiai koartikulációs jelenségek, ejtésvariációk tanulmányozásával több munka is foglalkozik, pl. [Gósy 98] [Vicsi & Szaszák 04] [Zsigri & Tóth+ 04]. Sajnos – a kutatócsoportunkhoz köthető publikációkat nem számítva – a konkrét felismerési alkalmazásokról, felismerési eredményekről ezek a cikkek vagy nem szólnak, vagy hiányzik az eredmények valamely referenciamódszerhez történő összevetése.

Kutatócsoportunk kezdetben kézzel, beszédhangszinten szegmentált beszédadatbázisokat alkalmazott az akusztikus modellek tanítására, azaz a tanítás során explicite figyelembe lettek véve a fonológiai koartikulációs jelenségek (Bábel adatbázis [Vicsi & Vig 98]). Ekkor természetesnek volt tekinthető a törekvésünk, hogy kapcsolt szavas, illetve folyamatos számfelismerés esetén ne csak szóbelsőben, de szóhatárokon is explicite modellezzük a hasonulási, egybeolvadási és egyéb fonológiai koartikulációs jelenségeket. Kezdeti sikereket értünk el, a kapcsolt szavas és folyamatos számfelismerési hálózatokba kézi erővel integrált fonológiai koartikulációk hatására számottevő felismerési pontosságjavulás adódott [B4, C9, C10].

A továbblépést számunkra a nagyobb méretű tanító-adatbázisok olyan automatizált feldolgozása jelentette, ahol a tanítószövegben a fonológiai koartikulációk nem kézi úton, hanem automatikusan mennek végbe. A célunk az volt, hogy az akkor csak izolált szavas felismerést megengedő teszthálózatban szóbelsejei fonológiai koartikulációk a tanítószövegben szóhatárokon is automatikusan végbemehessenek. Ennek lehetővé tételével sikerült szignifikánsan javítani a felismerési pontosságot városnévfelismerési feladatok esetén. Először környezetfüggetlen beszédhangmodellek mellett [J2], majd környezetfüggő beszédhangmodellek [B3] alkalmazásával is. A módszer azonban csak lineáris szövegek esetén volt használható, azaz tanítószövegek és izolált szavas tesztszövegek automatikus fonológiai átírására, de folyamatos beszédfelismeréshez használt összetett, elágazó hálózat kezelésére már nem. Továbbá, el kell ismernünk, hogy tanításnál, tesztelésnél következetes implicit fonológiai koartikulációmodellezési kísérleteket nem végeztünk.

Először [Szarvas & Furui 02] alkalmazott súlyozatlan FST alapú explicit fonológiai koartikulációs modelleket magyar nyelvű (mikrofonos) folyamatos beszéd felismerésére. Hasonlóan [Hazen & Hetherington+ 02]-hoz, 8.3% relatív felismerési hiba csökkenést ért el környezetfüggő beszédhangmodellek esetén. Azonban a tanító-adatbázis feldolgozási módjának nem ismertetése, a tesztadatbázis kis mérete és a szignifikancia vizsgálatok hiánya miatt az eredmények értékelése nehézségekbe ütközik.

A következőkben az általunk elérhető legnagyobb magyar nyelvű (telefon)beszéd adatbázisokon vizsgáljuk a fonológiai koartikuláció explicit modellezésének hatását, különböző feltételek mellett, folyamatos gépi beszédfelismerési alkalmazásokban. Az eredmények nagyobb részét korábban a [C6]-ban mutattuk be.

5.2. Fonológiai koartikulációk a magyar nyelvben

A modern nyelvtudomány a “kiejtési szabályok” néven összegyűjtött hasonulási, összeolvadási stb. jelenségeket fonológiai koartikulációs jelenségeknek hívja [Gósy 04]. Ezek főbb ismérve, hogy egy vagy több beszédhang fonémaértéke megváltozik a kiejtés során (pl. *azt → a sz t*). A megváltozás lehet összetettebb jelenség, beleértve a kiesést vagy betoldást is (pl. *értsd → é r dzs d, tea → t e j a*). Külön említendők a szóhatárokon fellépő fonológiai változások (pl. *értsd te → é r dzs d _ t e* vagy *é r cs t e*), melyek attól is függhetnek, hogy tart-e szünetet a beszélő a két szó között vagy sem, illetve, természetesen attól is, hogy milyen hanggal kezdődik a következő szó.

A fonológiai koartikulációs jelenségek egy lehetséges csoportosítása a következő:

- Zöngésségi (részleges és teljes) hasonulások: *adta → a tt a, lékbe → l é g b e*
- Képzés helye szerinti (részleges és teljes) hasonulások:
  azonban → a z o m b a n, önmaga → ö mm a g a
- Mássalhangzó-rövidülések: *állt → á l t*
- Összeolvadások: *látja → l á tty a, utca → u cc a, kétség → k é ccs é g*
- Egyéb kiesések, betoldások:
  parasztkolbász → p a r a sz k o l b á sz, tea → t e j a

5.3. A fonológiai koartikulációs jelenségek explicit modellezése

Mivel a fonológiai koartikuláció véges és kis elemszámú fonémák környezettől függő megváltozásait jelenti, jól modellezhető a beszédfelismerésben egyébként is alkalmazott véges állapotú átalakítókkal. Bár az egyes jelenségek megvalósulásaihoz elvileg valószínűségek is társíthatók, mi ezzel a lehetőséggel nem élve, a [Szarvas & Furui 02]-hoz hasonlóan súlyozatlan transzducereket alkalmaztunk.

A kísérletekben az először [J6]-ban bemutatott hierarchikus fonológiai koartikulációs szabályrendszer WFST megfelelőjét használtuk, melyet az alább részletezett módon állítottunk össze elemi szabálytípusoknak megfelelő véges átalakítókból.

P_1: Zöngésségi hasonulás /kötelező/

P_2: Összeolvadás + Rövidülés /kötelező/

P_3: Képzés helye, módja szerinti részleges hasonulások /opcionális/

P_4: Képzés helye, módja szerinti teljes hasonulások /opcionális/

A fonológiai koartikulációs modell, P, az alábbi kompozíciósorozattal adódik:

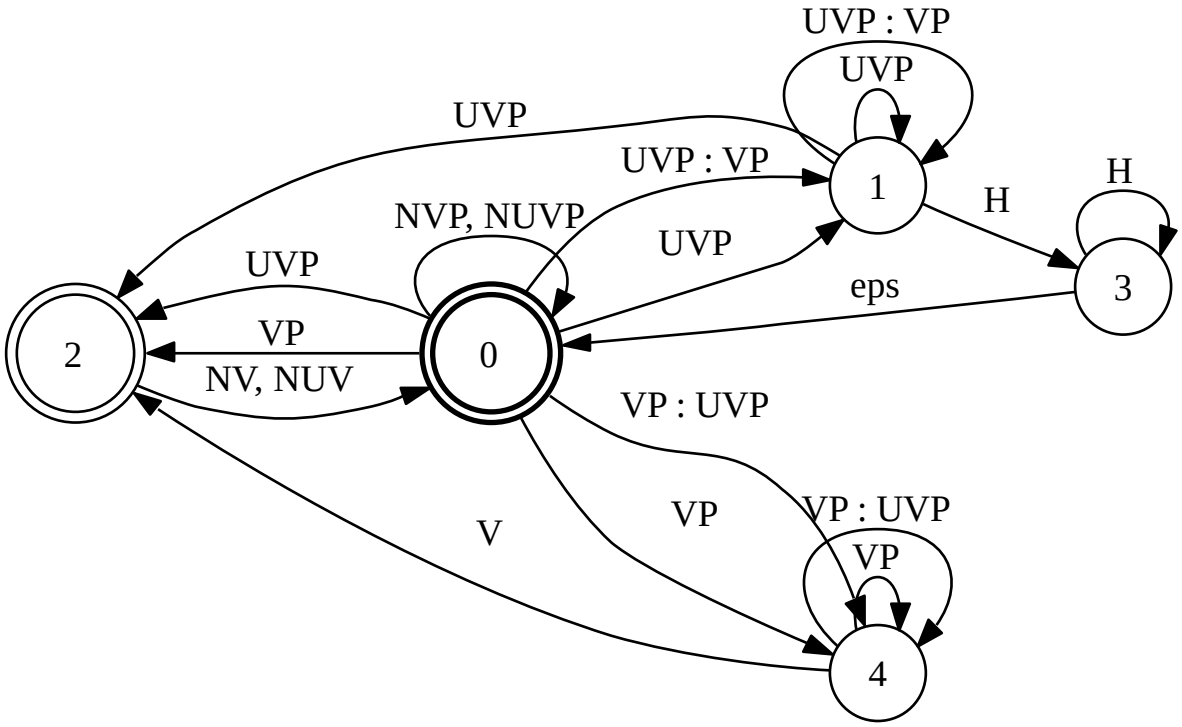
$$P = P2 \circ P4 \circ P3 \circ P2 \circ P1 \qquad (5.1)$$

Ez a modell az előző pontban említett fonológiai koartikulációs jelenségek közül mindegyiket explicit módon, *szóhatárokon átívelve* (is) kezeli. Kivételt csak az "egyéb kiesések, betoldások" képeznek, mert ezek esetlegesek, ritkák és automatizáltan nem állíthatók elő. Megjegyezzük, hogy a szóhatárokon átívelő koartikulációt csak akkor tesszük lehetővé, ha a két szó közé nem esik szünet a kiejtés során.

A rész fonológiai transzducerek a következők szerint kerültek kialakításra.

5.3.1. Zöngésségi hasonulás (P1)

A zöngésségi hasonulásokra az jellemző, hogy a későbbi hang hat vissza az előbb elhangzóra, így a folyamatban résztvevő mássalhangzókból gyakorlatilag az utolsó határozza meg zöngésségi jegyét az egész csoportra. Az ilyen jelenségeket újraíró szabályokkal lehet kezelni [Kaplan & Kay 94]. A konkrét WFST megvalósítást az 5.1. ábra szemlélteti.

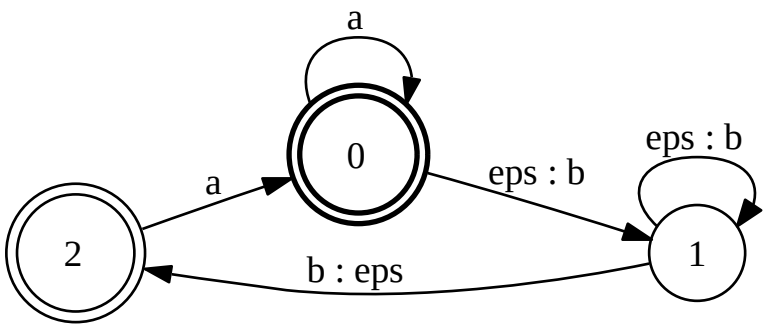


5.1. ábra. A magyar zöngésségi hasonulások kötelező, újraíró FST reprezentációjának szemléltetése. Jelmagyarázat – V: voicing /zöngésítő/, UV: unvoicing /zöngétlenítő/, VP: voiced pair /zöngés-zöngétlen pár előbbi fele/, UVP: unvoiced pair /zöngés-zöngétlen pár utóbbi fele/, N: negation /negáló jel/, H: [h] fonéma.

Fontos megjegyezni, hogy a beszédszünetet is fonéma értékűnek tekintjük, mely nem vesz részt a zöngésségi hasonulásban. Így ha két szó között a felismerésnél szünet adódik, a zöngésségi hasonulás szabályaink szerint nem terjedhet át az előző szóra. Ha viszont nincs szünet, akkor ugyanúgy végbemegy, mint szó belsejében.

5.3.2. Összeolvadás + rövidülés (P2)

Itt lényegében arról van szó, hogy amikor kettő, csak zöngésségi jegyben eltérő mássalhangzó kerül egymás mellé, a P1 hatására két egyforma fonéma keletkezhet, melyek helyett egy hosszú mássalhangzót kellene képezni. Viszont elképzelhető, hogy egy harmadik mássalhangzó mellett voltak, ilyenkor a hosszú mássalhangzót röviddel kell helyettesíteni. (Mivel az elemi akusztikai modellezésnél – a HMM keretrendszer korlátai miatt – nem tudunk jól időtartam-információt modellezni, a hosszú mássalhangzókat mindig röviddel helyettesítjük. Így a rövidülési jelenségekkel a konkrét megvalósításnál nem foglalkozunk.)



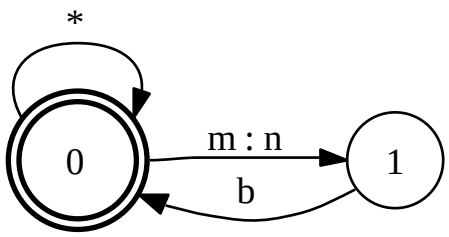
5.2. ábra. Újraíró kötelező szabály FST szemléltetése csak mássalhangzó rövidítésre, 2 betűs ABC-re.

5.3.3. Képzés helye, módja szerinti részleges hasonulások (P3)

Azokat a típusú hasonulásokat modelleztük itt, ahol két szomszédos mássalhangzó közül csak az első változik meg a koartikuláció során, a második változatlan marad. A kísérletekben az alábbi szabályokat implementáltuk:

1. msh	2.msh	1.msh’
n	p	m ;
n	b	m ;
n	ty	ny ;
n	gy	ny ;

Mivel ezek a hasonulások az artikulációtól függhetnek, opcionálisnak tekintettük érvényesülésüket.



5.3. ábra. Képzés helye, módja szerinti részleges hasonulások FST szemléltetése opcionális, nem újraíró szabályként.

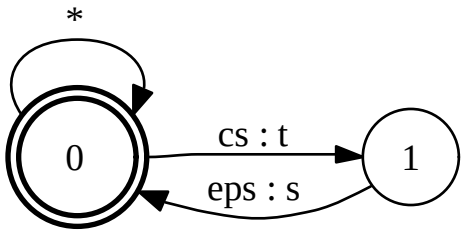
5.3.4. Képzés helye, módja szerinti teljes hasonulások (P4)

Ide azok a hasonulások kerültek, amelyeknél a két szomszédos mássalhangzó helyett egy harmadik mássalhangzót ejthetünk ki. Ezek a szabályok is opcionálisak, mert beszédstílustól és egyéb szempontoktól függően nem minden esetben valósulnak meg. Az alábbi szabályrendszert valósítottuk meg véges állapotú transzducerként.

1. msh	2. msh	1+2. msh
t	s	ccs ;
t	sz	cc ;
t	cs	ccs ;
t	c	cc ;
d	zs	ddzs ;
d	z	ddz ;
d	dzs	ddzs ;
d	dz	ddz ;
t	j	tty ;
d	j	ggy ;
n	j	nny ;
ty	j	tty ;
gy	j	ggy ;
ny	j	nny ;
t	ty	tty ;
d	gy	ggy ;
n	ny	nny ;

n	m	mm ;
l	j	jj ;
l	r	rr ;

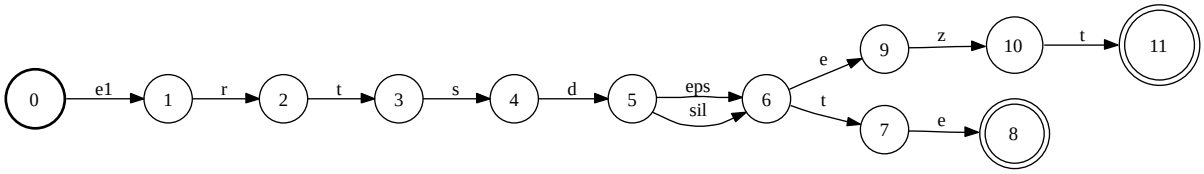
Látható az (5.1) szerinti szabályhierarchia hatékonysága: a fenti táblázatban alig negyedannyi szabályra van szükség, mintha a P1 transducert nem alkalmaztuk volna előzőleg.



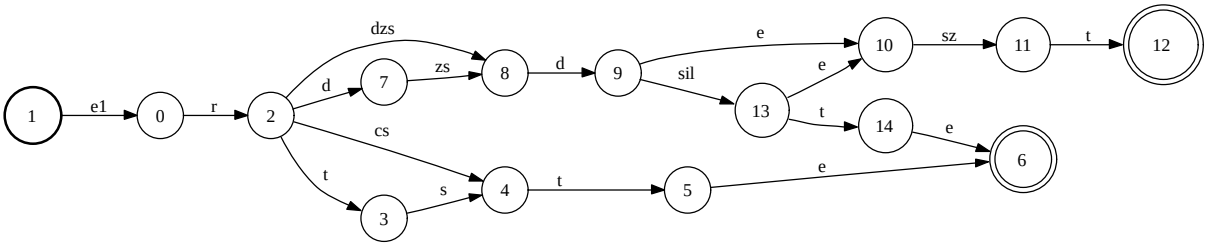
5.4. ábra. Képzés helye, módja szerinti teljes hasonulások FST szemléltetése opcionális, nem újraíró szabályként.

5.3.5. A P fonológiai véges átalakító hatásának szemléltetése

A P transzducer fonológiai koartikulációs modellezési képessége a következő példákkal szemléltethető:



5.5. ábra. Kapcsolt szavas fonémaszintű felismerési hálózat (F), jelöletlen fonológiai koartikulációval. (Az „értsd te” és „értsd ezt” szókapcsolatok nyers fonológiai szintű FST reprezentációja.)



5.6.ábra. Kapcsolt szavas fonémaszintű felismerési hálózat explicit fonológiai koartikukáció modellezésével (P o F). (Az „értsd te” és „értsd ezt” szókapcsolatok felszíni fonológiai szintű FST reprezentációja.)

5.3.6. Nem modellezett fonológiai koartikulációs jelenségek

Nem modelleztük többek között a hiátustöltést. Ennek oka, hogy a preferált alacsonyabb szintű (fonetikai) kiejtésmodellezés (trifón) véleményünk szerint teljes mértékben lefedi ezt a jelenséget. Nem modelleztünk továbbá olyan mássalhangzó-kiejtési jelenségeket, melyek nem jól algoritmizálhatók, inkább csak egyedileg írhatók le. Pl. ébresztget → ébrezget stb.

5.4. Gépi beszédfelismerési kísérletek

A fonológia koartikulációs modellezési megközelítések kiértékelésére az előző fejezetben leírtakhoz hasonló felismerési kísérleteket végeztünk. Mivel kisebb-nagyobb különbségek adódtak, a biztonság kedvéért újra közöljük a felismerési beállításokat, adatbázis paramétereket. Csak folyamatos beszéd felismerését célzó teszteredményeket közlünk, mivel azok a korábbi tapasztalatok alapján sokkal érzékenyebben mutatják ki a különböző modellezési megközelítések közti különbséget, mint az izolált szavas tesztek. A következő kísérletekben az előzőekben bemutatott explicit fonológiai koartikulációs modellezést vetjük össze az e nélküli, ill. implicit fonológiai koartikulációs modellezési megközelítéssel számos kísérleti konfigurációban.

5.4.1. Beszédadatbázisok

A tanító és tesztelő adatbázisokat – csakúgy, mint a 4. fejezetbeli kísérleteknél – a legnagyobb magyar telefonos beszédadatbázisok, az MTBA, a Besztel, a SpeechDat és a Tesztel összességéből [Vicsi & Tóth 02], [Vicsi et al.] alakítottuk ki. Ezek az adatbázisok elsősorban olvasott beszédet, valamint kisebb arányban spontán bemondásokat is tartalmaznak.[12]

Tanítóhalmazok: Tanítás céljára az MTBA, Besztel, és a SpeechDat adatbázis 500-400-400 beszélőjének azon felvételeit jelöltük ki, melyek nem „o”, és „z” jelzésűek, azaz nem tartalmaznak tulajdonneveket és bizonyos típusú mondatokat. A SpeechDat esetén csak egy szűkebb halmazt, a fonetikailag változatos szavakat és mondatokat (kivéve a „z” jelzésűeket) használtuk.

A teljes tanítóhalmaz mellett annak bizonyos részhalmazait is képeztük, hogy a különféle koartikulációs modellezési eljárások tanító-adatbázismérettől való függését is vizsgálhassuk. Sem a tanítóhalmazokban, sem a későbbi teszthalmazokban nem végeztünk szűrést az annotációnál zajosnak minősített felvételekre. Kizárólag azokat a felvételeket hagytuk ki, melyeknek az eleje vagy vége az annotáció szerint nem került rögzítésre.

A tanítóhalmazok jelölése és tartalma:

- **M**: Az MTBA fonetikailag változatos mondatai és szavai, 500 beszélő, 6000 felvétel
- **MM**: Az MTBA összes tanítófelvétele, 500 beszélő, 19000 felvétel.
- **MM_BS**: Az MTBA és a Besztel összes tanítófelvétele, 900 beszélő, 39000 felvétel.
- **MM_BS_SD**: Az MTBA, a Besztel és a SpeechDat tanítófelvételei, 1300 beszélő, 44000 felvétel.

A felismerési feladat: Az általános tapasztalat szerint a beszélőfüggetlen folyamatos beszédfelismerés támasztja a legnagyobb igényeket az alkalmazott modellekkel szemben. Ezért olyan *általános* folyamatos beszédfelismerési feladatot próbáltunk definiálni, ami a rendelkezésre álló adatbázisokkal megvalósítható. Természetesen adódott, hogy az adatbázisok azon mondatait tartalmazó bemondásokat ismertessük fel, melyek nem szerepelnek a tanítóhalmazokban. A beszélőfüggetlenség követelménye miatt azon felvételeket is ki kellett zárnunk, melyeknek a beszélőjét felhasználtuk a tanítás során.

[12] Bővebben lásd: 4.4.1.-et és a hivatkozott referenciákat [Vicsi & Tóth 02], [Vicsi et al.].

Teszthalmazok: A teszthalmazokat tehát úgy állítottuk össze, ne legyen átfedés a tanítóhalmazban szereplő beszélőkkel. Így a tanításnál fel nem használt 220 beszélőtől (Besztel 100, SpeechDat 100, TeszTel 20) kerültek felvételek a teszthalmazokba. Összesen 2385 felvételt kaptunk, melyeket a tanító-adatbázishoz való illeszkedés mértéke szerint két halmazra bontottunk.

A folyamatos beszédfelismerésnél fonológiai és nyelvi illeszkedésről is beszélhetünk. Az egyik halmazba azokat a mondatokat válogattuk, amelyeknek *szöveges tartalma* egyezett az akusztikus modelltanításnál használt mondatokéval (fonológiai illeszkedés), valamint amelyeknek szöveges tartalma a nyelvi modell tanításakor is felhasználásra került (nyelvi illeszkedés), ez az „illeszkedő" ("Matched": M) halmaz. A másik, „nem illeszkedő" ("Unmatched": U) teszthalmazba azok a felvételek kerültek, melyek szövegtartalma sem az akusztikus, sem a nyelvi modell tanításkor nem lett felhasználva. Egyéb halmazt nem vizsgáltunk.

A teszthalmazok jelölése és tartalma:

- **M**: Nyelvi és fonológiai szempontból a tanításhoz illeszkedő mondatok, 220 beszélő, 1973 felvétel: „s" jelzésű mondatok a Besztel-ből és a SpeechDat-ból, „s1" és „s2" jelzésű mondatok a TeszTel-ből.
- **U**: Sem nyelvi és sem fonológiai szempontból a tanításhoz nem illeszkedő mondatok, 220 beszélő, 412 felvétel: „z" jelzésű mondatok a Besztel-ből és a SpeechDat-ból, „s3" jelzésű mondatok a TeszTel-ből.

5.4.2. Beszédfelismerési paraméterek, beállítások

Lényegkiemelés: Lényegkiemelési paraméterekként a bemenő beszédjelből MFCC (Mel Frequency Cepstral Coefficients) 12 dimenziós vektorokat képeztünk, melyekhez *logE* (keretenkénti logaritmikus energia) paramétert is csatoltunk, majd dinamikus Delta és Delta-Delta értékeket számítottunk. A statikus energiát végül kicsatolva összesen 38 dimenziós jellemzővektorokat kaptunk. Mind a tanítás, mind a tesztelés során alkalmaztuk a vak csatornakiegyenlítés (Blind Equalization) módszerét [C14].

Elemi akusztikus modellek: Az atomi modellek rejtett Markov-modell állapotok voltak rögzített hurok és továbblépési valószínűségekkel. Állapotonként maximum 10 Gauss függvényből álló folyamatos megfigyelési sűrűségfüggvényeket használtunk [Young 06].

Fonetikai koartikulációs modellek (H_{mono} és H_{tri} o C): Mind a monofón mind a trifón modelleknél a beszédhangokat 3 elemi akusztikus modellre képeztük le, az előbbi esetben a környezettől függetlenül az utóbbi esetben az ML döntési fa alapján a fonetikus környezettől függően [Young 06]. A döntési fákat - és így a H_{tri} leképezést - tanítóhalmazonként és fonológiai koartikulációs modellenként újraépítettük. Szóhatárokon átívelő trifón modellezést használtunk.

Fonológiai koartikulációs modell (P): A 5.3. alfejezetben ismertetett módon állítottuk össze a „kiejtési szabályok" néven közismert fonológiai koartikulációs jelenségek nagy részét modellező véges állapotú átalakítót. A modell a szóhatárokon átívelve is modellezi a koartikulációt.

Szótármodell (L): A kiejtési modellek nyers, fonológiai koartikulációkat nem tartalmazó fonemikus átiratait automatikusan állítottuk elő [J2]. Allofónikus változatokat nem jelöltünk, továbbá a hosszú és rövid mássalhangzókat sem különböztettük meg. Így – a szünetmodelleket nem számítva – összesen 39 fonológiai kategóriát használtunk. A szünetmodell háromállapotú környezetfüggetlen modell volt.

Az alkalmazott 5561 elemű szótár az összes előforduló szót tartalmazta (beleértve az illeszkedő és a nem illeszkedő teszthalmaz szavait), így szótáron kívüli elemek kezelésére nem volt szükség.

Nyelvi modell (G): A folyamatos felismerésnél szó-trigram nyelvi modelleket alkalmaztunk Katz-féle visszametszéssel [Katz 87] és Good-Turing valószínűség-újraelosztással [Good 53]. A tanítószöveg az illeszkedő tesztmondatok szövege alapján készült úgy, hogy minden különböző mondatot csak *egyszer* szerepeltettünk. Így az illeszkedő mondatokon PP=40-es perplexitást, a nem illeszkedő tesztmondatokon PP=6230-as (nagyon magas, azaz igen kedvezőtlen) perplexitás értéket kaptunk. A nyelvi modellezésre az SRILM eszközt alkalmaztuk [Stolcke 02].

A beszédfelismerési tudásforrások integrációját és optimalizációját a 2. fejezetben ismertetett WFST módszerekkel végeztük az AT&T FSM Toolkit segítségével. A felismerő motor a [C7]-ban említett, optimalizált Viterbi-algoritmuson (dinamikus programozáson) alapuló eszköz volt. A keresési mélységet fixen úgy állítottuk be, hogy a felismerési pontosság minden esetben bőven a telítési szakaszra essen, így különböző RTF mellett is összehasonlíthatóak az eredmények.

5.4.3. A fonológiai koartikulációs modellek kiértékelése kézi, fonológiai szintű tanító-adatbázis-feldolgozás mellett

Itt azt vizsgáltuk, hogy ha a tanító-adatbázis eredeti kézzel ellenőrzött fonológiai átiratait és szegmentációját használjuk a beszédhangmodelleket tanításához, a tesztelésnél – ahol ezeket a fonológiai kiejtési modelleket gépi úton állítjuk elő – van e jelentősége, és mekkora a fonológiai koartikulációs jelenségek explicit modellezésének.

A vizsgálatra egyedül az M-jelű tanítóhalmaz volt alkalmas (MTBA, fonetikailag változatos szavak, mondatok).

A kiértékelést explicit *fonetikai* koartikulációs modellezés – azaz döntési fa alapján állapotcsoportosított trifón modellek – mellett végeztük, mivel korábbi vizsgálataink szerint (lásd a 4.4. alfejezetet) ez jelentette a nem vizsgált paraméterek optimális beállítását.

Az alábbi két felismerési hálózattal végeztünk kísérleteket:

- H_{tri} o C o L o G – nincs fonológiai koartikuláció-modellezés
- H_{tri} o C o P o L o G – explicit fonológiai koartikuláció-modellezés

Mivel a beszédhangmodelleket kézzel ellenőrzött – tehát a fonológiai koartikulációkat jelölő – fonetikus szegmentáció mellett tanítottuk, azok alapvetően nem modellezték még implicite sem a fonológiai koartikulációs jelenségeket. Így a P alkalmazásától szignifikáns javulást vártunk. Az eredményeket az 5.1. a) és b) táblázat mutatja.

5.1. a) és b) Táblázat. Az illeszkedő (***M***) és nem illeszkedő (***U***) teszthalmazok folyamatos beszédfelismerési eredményei explicit fonológiai koartikulációs modell nélkül, illetve annak alkalmazása mellett, kézi tanító-adatbázis-feldolgozás esetén [%]-ban.

a) ***M*** *(PP = 40)*	*Corr.*	*Acc.*
H_{tri} o C o L o G	93.05	91.40
H_{tri} o C o P o L o G	93.99	92.57
Relatív javulás	*13.6*	

b) ***U*** *(PP = 6230)*	*Corr.*	*Acc.*
H_{tri} o C o L o G	60.84	49.45
H_{tri} o C o P o L o G	62.02	51.09
Relatív javulás	*3.2*	

Az illeszkedő teszthalmaz esetén kétszámjegyű relatív hibacsökkenés figyelhető meg, ugyanakkor a nem illeszkedő halmaz esetén a relatív javulás szerényebb. Az szófelismerési hibacsökkenés szignifikanciáját az 1.6.2. szerinti *2 mintás Wilcoxon-próba* segítségével ellenőriztük. Standard 5% szignifikancia-szint mellett (95% konfidencia szint) mindkét teszthalmaz esetén *szignifikáns* javulást tapasztaltunk.

Az M-hez képest az U teszthalmazon – ugyanazon felismerési feladatban – mért sokkal gyengébb felismerési eredményeket a vonatkozó igen magas nyelvi modell perplexitás (PP) magyarázza.

5.4.4. A fonológiai koartikulációs modellek kiértékelése következetes tanító-adatbázis-feldolgozás mellett

Az előző vizsgálatnál a referencia rendszerben egyáltalán nem modelleztük a fonológiai koartikulációs jelenségeket, mégis csak kismértékű – bár szignifikáns – javulást kaptunk az explicit modell alkalmazásával. Ezért felmerült a kérdés, hogy következetes gépi szegmentációt alkalmazva és nagyobb tanító-adatbázisokat használva is tapasztalható-e érdemi felismerési hiba csökkenés a P véges átalakítónak köszönhetően.

A következő gépi fonetikus szegmentációs módszert dolgoztunk ki a következetes fonológiai modellezés érdekében. A legnagyobb tanítóhalmazra (MM_BS_SD) képeztük a lineáris G_{train} „nyelvi modellt”, majd előállítottuk a *tanítóadatokra* vonatkozó felismerési hálózatokat:

- H_{tri} o CD o L o G_{train} – *implicit* fonológiai koartikuláció-modellezés
- H_{tri} o CD o P o L o G_{train} – explicit fonológiai koartikuláció-modellezés

Kezdeti beszédhangmodelleket tanítottunk be az M tanító halmaz manuális szegmentációja alapján. Ezekkel kényszerített felismerést („forced alignment”) végezve megkaptuk a fonológiai koartikulációt implicit valamint explicit módon tartalmazó gépi fonetikus szegmentációkat.

A különböző tanítóhalmazok és felismerési hálózatok esetén mindig a megfelelő tanítású beszédhangmodelleket alkalmaztuk. Összesen tehát 4 X 2 akusztikus modell halmazt vizsgáltunk 2 felismerési hálózattal.

A tesztelésnél használt felismerési hálózatok az előzőekben vizsgáltakkal azonosak voltak:

- H_{tri} o CD o L o G – *implicit* fonológiai koartikuláció-modellezés
- H_{tri} o CD o P o L o G – explicit fonológiai koartikuláció-modellezés

Fontos megjegyezni, hogy a következetes tanítás és tesztelés miatt a P kihagyása már nem jelenti azt, hogy a fonológiai koartikulációt egyáltalán nem, hanem, hogy implicite, vagyis alacsonyabb, azaz beszédhang szinten modellezzük.

5.2. a) és b) Táblázat. Az illeszkedő és nem illeszkedő teszthalmazok folyamatos beszédfelismerési eredményei következetes gépi adatbázis-feldolgozás mellett [%]-ban.

a) ***M*** *(PP = 40)*	*M*		*MM*		*MM_BS*		*MM_BS_SD*	
	Corr.	*Acc.*	*Corr.*	*Acc.*	*Corr.*	*Acc.*	*Corr.*	*Acc.*
H_{tri} o C o L o G	94.13	92.54	93.82	91.97	94.22	92.55	94.47	92.93
H_{tri} o C o P o L o G	94.24	92.69	93.41	91.66	94.14	92.54	94.78	93.05
Relatív javulás	2.0		-3.9		-0.1		1.7	

b) ***U*** *(PP = 6230)*	*M*		*MM*		*MM_BS*		*MM_BS_SD*	
	Corr.	*Acc.*	*Corr.*	*Acc.*	*Corr.*	*Acc.*	*Corr.*	*Acc.*
H_{tri} o C o L o G	61.95	48.16	61.27	47.66	64.24	52.34	64.42	52.02
H_{tri} o C o P o L o G	62.34	50.14	61.24	48.20	64.09	51.91	65.13	53.20
Relatív javulás	3.8		1.0		-0.9		2.5	

Ahogy az 5.2 a) és b) táblázatok mutatják, az implicit és explicit fonológiai koartikulációs modellek beszédfelismerési eredményei között a különbség minimális. A szóhibaarányokon Wilcoxon-próbával végzett szignifikancia-vizsgálatok *egyetlen esetben sem mutattak ki szignifikáns javulást* a *p=0.05*-ös szinten, sőt a hiba nem is csökkent minden esetben.

Észrevehető, hogy a kézi helyett gépi feldolgozás (fonológiai átírás és szegmentáció) az M halmaz esetében érdemben nem rontott az eredményeken. Tehát következetes gépi tanitóadatbázis-feldolgozás mellett gyakorlatilag ugyanolyan jó eredmények érhetők el fonológiai modell *nélkül* is, mint a manuálisan szegmentált adatbázissal *és* explicit, szóhatárokon átívelő hasonulási modellel.

5.3. táblázat. Az folyamatos gépi beszédfelismerés átlagos számításigénye (Real-Time Factor).

RTF:	*M*	*MM*	*MM_BS*	*MM_BS_SD*
H_{tri} o C o L o G	0,57	0,69	0,78	0,85
H_{tri} o C o P o L o G	0,64	0,77	0,88	0,92

Ahogy az 5.3. táblázat mutatja, az explicit modellezés – attól függetlenül, hogy nem hozott érdemi javulást a felismerési eredményekben – lassította a felismerési folyamatot.

5.4.5. A fonológiai koartikulációs modellek kiértékelése környezetfüggetlen beszédhangmodellezés mellett

Mivel ez idáig a *fonetikai* koartikulációt explicit módon modelleztük a kísérleti rendszerekben nem tudtuk közvetlenül megvizsgálni, hogy önmagában a fonológiai koartikuláció explicit modellezése a fonetikai *nélkül* mennyiben javítja a gépi beszédfelismerés pontosságát. Noha az eddigi tapasztalatok alapján jelentős javulás nem volt várható, ellenőrzésképpen elvégeztük az alábbi, „monofón" modellekkel készült kísérleteket.

Csak a legkisebb (M) és legnagyobb (MM_BS_SD) tanítóhalmazzal tanítottunk, és következetes gépi adatbázis-feldolgozást (fonológiai átirat és szegmentáció) alkalmaztunk.

A tesztelésnél használt felismerési hálózatok az következők voltak:

- H_{mono} o L o G – *implicit* fonológiai és fonetikai koartikuláció-modellezés
- H_{mono} o P o L o G – explicit fonológiai és implicit *fonetikai* koartikuláció-modellezés

5.4. a) és b) Táblázat. Az illeszkedő és nem illeszkedő teszthalmazok folyamatos beszédfelismerési eredményei következetes gépi adatbázis-feldolgozás és környezetfüggetlen beszédhangmodellezés mellett [%]-ban.

a) ***M*** *(PP = 40)*	*M*		*MM_BS_SD*	
	Corr.	*Acc.*	*Corr.*	*Acc.*
H_{mono} o L o G	85,34	84,32	80,74	79,52
H_{mono} o P o L o G	86,26	85,12	82,74	81,44
Relatív javulás	*5,1*		*9,4*	

b) ***U*** *(PP = 6230)*	*M*		*MM_BS_SD*	
	Corr.	*Acc.*	*Corr.*	*Acc.*
H_{mono} o L o G	29,22	23,00	25,58	20,54
H_{mono} o P o L o G	29,87	23, 65	26,65	21,45
Relatív javulás	0,9		1,5	

Amint a 5.4. táblázat mutatja, a fonológiai koartikuláció explicit modellezése javít ugyan valamennyit a felismerési eredményeken, azonban ennek mértéke összehasonlíthatatlanul kisebb, mint amennyit az explicit *fonetikai* koartikuláció-modellezéssel sikerült elérni (lásd 4. fejezet).

5.5. Összefoglalás

Megvalósítottuk a magyar nyelvi fonológiai koartikulációs jelenségek jelentős részét explicit módon modellező *P* transzducert. Ennek beszédfelismerési alkalmazása révén bizonyos szuboptimális beállítások mellett szignifikáns javulás volt tapasztalható. Az explicit fonológiai koartikuláció modellezés általi javulás azonban eltűnt, ha környezetfüggő beszédhangmodelleket használtunk és következetes fonológiai megközelítést alkalmaztunk tanításkor és teszteléskor. Tekintve, hogy az abszolút felismerési pontosságok az utóbbi beállításban voltak a legjobbak, a *P* alkalmazását elvetjük, hiszen mind a rendszer komplexitását, mint a felismerési időt a nélkül növeli, hogy érdemben a javítana a felismerési eredményeken. Ez a felismerésünk egybevág [Jurafsky & Ward+ 01]-éval, amit ők más nyelvre, más módszerekkel kaptak. Így egyéb nyelvekre is igaz lehet a megállapítás, hogy a

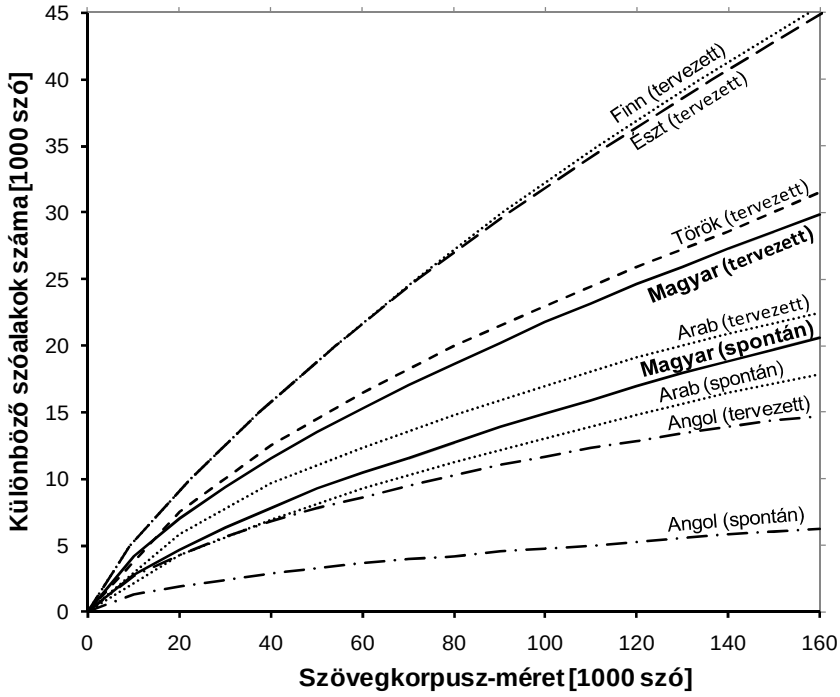
statisztikai gépi beszédfelismerésben következetes tanítás-tesztelés és explicit környezetfüggő beszédhang modellek használata mellett nem feltétlenül szükséges a fonológiai koartikuláció explicit, (súlyozatlan) szabály alapú fonológiai szintű modellezése.

6. Lexikai modellezés spontán magyar nyelvű beszéd gépi felismeréséhez

Magától értetődő, hogy a beszéd akusztikai modellezésénél a szavakat hangokra bontjuk. Nem ilyen egyértelmű viszont, hogy a (nagyszótáras) folyamatos beszéd felismeréséhez alkalmazott nyelvi modellezésnél is szükséges-e a szavakat kisebb egységekre bontani, és ha igen, hogyan. Ezzel a kérdéssel a lexikai modellezés foglalkozik. A szavaknál kisebb lexikai egységek használata mind a nyelvi modellek, mind a kiejtési modellek alapegységeit meghatározza. A lexikai modellezés célja olyan lexikon és technika előállítása, mely a morfológiailag változatos nyelveknél (is) lehetővé teszi a szótárméret, illetve a nyelvi modellezés adatelégtelenségi („data sparsity") problémáinak kezelését, kézben tartását.

6.1. Bevezetés

A magyar nyelv és a hozzá hasonló „morfológiailag gazdag" nyelvek szó alapú modellezése több problémát is felvet, mert a ragozás, a tőváltás és a szóösszetétel miatt rendkívül nagy számú szóalak (akár több millió) állhat elő már viszonylag kis méretű tanítókorpuszok esetén is. A nagy szóalaki változatosság azonban nemcsak nagy szótárméretet jelent, hanem emellett a szótáron kívüli (OOV: Out Of Vocabulary) szavak aránya is igen magas lehet. A legnagyobb problémát azonban – véleményünk szerint – a rosszul becsülhető nyelvi modell paraméterek okozzák. Ezért mindenképpen indokolt folyamatos magyar nyelvű beszédfelismerésnél a lexikai modellezés vizsgálata.



6.1. ábra. Spontán magyar (MALACH) beszéd szótárméret-növekedési görbéjének elhelyezése nemzetközi környezetbe [Creutz & Hirsimäki+ 07] és [C1] alapján.

Ahogy a 6.1. ábra mutatja, az adott méretű szövegkorpuszban megtalálható különböző szóalakok száma nagymértékben különbözik nemcsak nyelvenként, de beszédstílusonként is. Látható, hogy spontán beszéd esetén jellemzően kisebb a szóalaki változatosság, mint az adott nyelvű tervezett beszédnél, ami a morfológiai modellezés szempontjából kedvező.

Ugyanakkor a következőkben részletezendő nemzetközi tapasztalatok szerint éppen a relatíve kisebb szóalaki változatosság miatt kevésbé hatékonyak a nagy morfológiai változatosságú beszédfelismerési feladatokra kidolgozott módszerek a spontán beszédre. Valamint, a spontán beszéd kezelése további kiejtésmodellezési problémákat is felvet (lásd 7. fejezet.)

A következőkben röviden áttekintjük a nemzetközi és hazai lexikai modellezési módszereket, melyeket a morfológiai változatosság kezelésére próbáltak alkalmazni.

6.1.1. A morfológiailag gazdag nyelvek lexikai modellezési módszereinek áttekintése

Az elmúlt években számos cikk jelent meg, mely morfológiailag változatos nyelvek (pl. német, holland, arab, finn stb.) esetén szónál kisebb lexikai egységek használatát javasolja – elsősorban a szótáron kívüli szavak kezelését, illetve a szótárméret csökkentését segítendő.

A szóösszetételt kedvelő germán nyelvekre (német, holland, svéd) publikált megközelítések csak az összetett szavak feldarabolását célozzák, pl. [Berton & Fetter+ 96], [Larson & Willett+ 00], [Ordelman & Hessen+ 03]. Ezek a technikák azonban a felismerés pontosságát érdemben nem tudták javítani.

Ragozó nyelvek esetében – mint amilyen a finn, észt, török, és koreai – az uralkodó megközelítés a szavak felbontása előtagokra, tövekre és ragokra, azaz morfémaszerű lexikai egységekre [Hirsimäki & Creutz+ 06], [Puurula & Kurimo 07], [Arisoy & Can+ 09], [Kwon & Park 03]. Fontos kérdés, hogy milyen technikával történjen a szavak kisebb lexikai egységekre történő darabolása. Természetesen adódik a nyelvi szabályokon és adatázisokon alapuló morfológiai elemzők használata [Puurula & Kurimo 07], [Arisoy & Can+ 09], de a matematikai–statisztikai elvű szószegmentációs eljárások is egyre inkább tért hódítanak [Hirsimäki & Creutz+ 06], [Arisoy & Can+ 09], [Creutz & Hirsimäki+ 07]. A továbbiakban az egyszerűség kevéért minden szónál kisebb egységet morfnak fogunk nevezni, tekintet nélkül arra, hogy hogyan származtatták őket, és van-e önálló jelentésük.

Az egyik legsikeresebb, legjobb eredményeket elérő szó→morf szegmentálási módszer meglepő módon egy felügyelet nélküli, tisztán statisztikai eljárás(család), melyet Morfessor-nak hívnak. Az alapeljárás az MDL (Minimum Description Length) alapelvre [Risannen 78] épül, és több variációja van, melyekkel a morf-szótár mérete és jellege kontrollálható [Creutz & Lagus 05a], [Creutz & Lagus 05b]. Ezzel a megközelítéssel a finn és a török nyelv esetén nemcsak, hogy a szó lexikai modellek eredményét szignifikánsan meghaladó felismerési pontosságot értek el, de számos nyelvi szakértői tudást alkalmazó grammatikai morf alapú megközelítés teljesítményét is sikerült túlszárnyalni [Hirsimäki & Creutz+ 06], [Arisoy & Can+ 09]. Észt nyelv esetén is jó, a nyelvtani morfokét megközelítő eredményt sikerült elérni [Puurula & Kurimo 07].

Ugyanakkor meg kell jegyeznünk, hogy minden eddigi eredmény tervezett beszédre vonatkozott, és a korábbról ismeretes egyetlen eredmény spontán nyelvű beszéd lexikai modellezésénél negatív volt [Creutz & Hirsimäki+ 07] (ECA: Egyptian Colloquial Arabic esetében). Morfológiai tudás hozzáadásával csak szerény mértékben sikerült az ECA beszédfelismerési feladat eredményén javítani az ún. faktorált nyelvi modellel [Kirchoff & Vergyri+ 06], de a lexikai egységek szavak voltak. Hasonlóan adódott a cseh spontán nyelvű beszédfelismerésnél [Shafran & Hall 06]. Megjegyzendő még, hogy [Afify & Sarikaya+ 06] iraki arab nyelvű beszéd esetén sikeresen és érdemben javított morf alapú lexikai modellek

alkalmazásával, azonban a beszédstílus maga nem volt definiálva, csak annyit tudni, hogy a rendszer alapvetően rövid, tőmondatok fordítására szolgáló alkalmazáshoz készült.

Végül meg kell említenünk, hogy egyes nyelveknél sikerrel alkalmaztak egy speciális lexikai modellezési technikát „flat hybrid" néven, ahol az egyes ritka szavakat közvetlenül betűkre darabolták, míg a többi szót egyben hagyták [Bisani & Ney 05]. Ez az eljárás azonban a morfológiailag változatosabb nyelvek esetén nem váltotta be a hozzá fűzött reményeket [Puurula & Kurimo 07].

6.1.2. A magyar nyelvű lexikai modellezési megközelítések áttekintése

Az első, úttörő kísérlet a nagyszótáras magyar nyelvű beszéd folyamatos beszédfelismerésére a [Szarvas & Furui 03a], [Szarvas & Furui 03b] és [Szarvas 03]-ben részletezett módon történt. Itt szabály alapú morfológiai szegmentáció és analízis alapján morféma alapú beszédfelismerő készült olvasott sajtószöveg felismerését célozva. A szerző morfo-szintaktikai szabályok hozzáadásával javított a beszédfelismerés hatásfokán, azonban közvetlenül nem vetette össze megoldását a szó alapú lexikai modellre épülő beszédfelismeréssel.

Magyar nyelvű radiológiai leletező alkalmazásban más kutatócsoport is publikált morféma-alapú eredményeket [Vicsi & Velkei+ 05], azonban náluk a nyelvi elemzővel [Prószéky & Tihanyi 93] előállított morféma egységekből a szavak visszaállítása nem történt meg, és így szóhibaarány-számítás sem történhetett morféma alapon. [Bánhalmi & Kocsor+ 05] szó lexikai egységek használata mellett morfo-szintaktikai jegyek hozzáadásával javított az orvosi diktáló rendszer pontosságán.

Bár [Szarvas 03] meggyőzően – igaz csak közvetve, lefedettségi mutatókkal – érvel a morféma alapú lexikai modellezés mellett, a mi esetünkben újságszöveg statisztikák alapján nem dönthető el, hogy spontán beszéd felismerésére milyen lexikai modell használata célszerű. Ezért többféle, szabály és statisztikai alapú megközelítés vizsgálata mellett döntöttünk. A kísérleteket a beszédadatbázis bemutatása után ismertetjük.

6.2. A MALACH spontán magyar nyelvű beszédadatbázis

A MALACH projekt célja, hogy a nagy mennyiségű beszélt nyelvű információt tartalmazó többnyelvű audio(vizuális) adatbázisokban a keresést, célzott információ-hozzáférést korszerű beszédtechnológiai eszközökkel segítse [Byrne & Doermann+ 04]. Esetünkben a spontán, nagyszótáras folyamatos gépi beszédfelismerés pontosságának javítása a feladat – magyar nyelven. A kísérletekben felismerendő beszédként Holokauszt-túlélők visszaemlékezéseinek rögzített hanganyaga szerepel. A teljes magyar nyelvű korpusz több mint 2000 óra hosszú (beszélőnként 2-3 óra). Konkrét célunk a korpusz egy kisebb, kézi erővel lejegyzett része alapján tanított gépi beszédfelismerő rendszer készítése és tesztelése volt. A projekt során 34 óra szövegátirata készült el. A továbbiakban ezt a transzkripcióval ellátott részt értjük a magyar MALACH adatbázis alatt.

A következőkben röviden ismertetjük az adatbázist, majd az alkalmazott lexikai és egyéb beszédfelismerési modelleket, illetve a különféle kísérleteket és eredményeiket. Mind a MALACH adatbázis, mind a kísérletek bővebb leírása megtalálható a [C1]-ben.

6.2.1. Beszéd- és szövegkorpusz jellemzők

A *hanganyag* 44.1 kHz-es frekvenciával mintavételezett, és általában az interjúalanyok otthonában készült. A téma természetéből adódóan a MALACH adatbázis beszélői általában idősek, beszédük néha inkoherens, megakadásokkal tarkított és esetenként erős angol befolyás érezhető a kiejtésen. Ugyanakkor a beszélők közül jópáran kivételesen tisztán, tagoltan a „standard”-hoz közeli módon artikuláltak.

A *szöveges lejegyzés* során az ortografikus és fonemikus variánsok egyaránt rögzítésre kerültek, amennyiben az annotáló úgy érezte, hogy az ortografikus alakból a kiejtett alak nem származtatható közvetlenül. A fonológiai koartikulációkat – az 5. fejezetben megírt tapasztalatok alapján – nem jelöltük, de az idegen szavak, betűzések, megakadások explicit módon meg lettek jelölve.

Az adatbázis egyes jellegzetességeit a 6.1. táblázatban foglaltuk össze. A MALACH adatbázis szótárméret-növekedési görbéjét pedig a 6.1. ábrán hasonlítottuk össze az angol nyelvű spontán és tervezett, illetve az egyéb nyelvű tervezett (pl. hírfelolvasási) beszéd görbéivel.

6.1. táblázat. A MALACH spontán magyar nyelvű adatbázis lejegyzett szövegének jellemzői

		Tanító-halmaz	Teszt halmaz	
			Gyengén illeszkedő	Illeszkedő
OOV arány [%]		–	14.6	14.2
Szó perplexitás		–	666.9	628.5
Előfordulási számok	Szavak	160k	16.8k	17.2k
	Betűk	810k	105k	108k
	Tört szavak	3496	435	207
	Értelmetlen szavak	1836	344	218
	Pongyola szavak	4528	647	522
	Betűzések	36	241	8
	Tulajdonnevek	3889	706	610
	Idegen szavak	422	137	60

6.2.2. Tanító- és teszthalmazok

Akusztikus és nyelvi modell tanítás céljára 104 beszélőtől személyenként 15 perces interjúrészleteket dolgoztunk fel, összesen 26 órányi beszédet. Az interjúrészletek a 30. percnél kezdődnek. Tesztelési célokra 8 órányi anyagot dolgoztunk fel, összesen 10 új beszélőtől, változó hosszakkal. A tanítóhalmazban összesen kb. 20 000 féle szóalak szerepelt, melyekből képzett szótárral a teszthalmaz szavainak kb. 15%-a nem fedhető le (OOV – „Out Of Vocabulary” – arány).

A teszthalmazt két részhalmazra bontottuk. Illeszkedő részhalmazként definiáltuk azon tesztfelvételeket, melyek a beszélgetések 30. percétől későbbi részleteket tartalmaznak. Ez a tesztfelvételek mintegy felét teszi ki. *Gyengén illeszkedő* részhalmaznak az előző komplementerét tekintjük, vagyis az első 30 percből származó felvételeket, ugyanis az interjú ezen fázisában rengeteg tulajdonnév, megakadás és betűzés hangozhat el (emiatt is maradtak ki a tanító-adatbázisból).

6.3. Az alkalmazott lexikai modellezési megközelítések

A fő feladat a megfelelő lexikai alapegységek meghatározása volt. Ezen alapegységek használandók a nyelvi modellezés során, valamint ezeket kell leképezni fonéma(szerű) akusztikai alapegységekké a kiejtésmodellezés során.

Jelen értekezésben a következő lexikai modellezési megközelítéseket alkalmazzuk és hasonlítjuk össze a velük elért beszédfelismerési pontosság szerint.

6.3.1. Szó alapú lexikai modellezés

Hagyományosan a gépi beszédfelismerés lexikai egységei a szavak. A MALACH projektben is – mind az angol [Ramabhadran & Juang+ 03], mind a morfológiailag változatosabb nyelvek esetén is (cseh, szlovák, orosz) – szó lexikai modelleket alkalmaztak [Psutka & Ircing+ 05]. Ezért a munkánk során ez a megközelítés természetes referenciaként szolgált.

6.3.2. Hunmorph – morfológiai adatbázis és szabályrendszer alapú morf szegmentálási módszerek

A Hunmorph morfológiai elemző [Trón & Németh+ 05] az utóbbi időkben egészült ki morf szegmentálási képességgel. Maga az elemző eszköz (Ocamorph) nyelvfüggetlen, azonban az elemzéshez, szegmentáláshoz felhasznált adatbázis (MorphDB.hu) [Trón & Halácsy+ 06] természetszerűleg nyelvfüggő. Kétféle szegmentálási mód tűnt az előzetes tesztek és elvárások szerint érdemesnek arra, hogy megvizsgáljuk őket a magyar MALACH felismerési feladatban:

- ***HSF (Hunmorph Strict Fallback)****:* a [Halácsy 06]-ban leírt alapelvhez hasonlóan a működés a következők szerint, szavankénti elemzéssel történik. Első körben a szegmentálandó szót nem összetett szóként kezeljük és keressük a lehetséges elemzéseknek megfelelő szegmentálásokat. Amennyiben létezik egy vagy több morfo-szintaktikai szabálynak megfelelő szegmentálás, az első legtöbb darabot eredményezőt választjuk, és az eljárás véget ér. Ha nem volt érvényes analízis, akkor összetett szóként tekintünk a bemenetre és az előzőekhez hasonlóan járunk el. Ha így sem volt sikeres elemzés, akkor a „guess" üzemmódba kapcsolva teszünk ugyanúgy, mint az előző két körben. Ilyenkor már minden esetben van érvényes szegmentáció. A kísérletek során a szótárelemek 93%-a az első körben, 4.3%-a a másodikban és a maradék 2.7% a harmadik körben lett felszegmentálva.

- ***HCG (Hunmorph Compound Guessing)****:* az előző módszerrel ellentétben az elemzés egyetlen körben megtörténik, amikor is az összetett szónak tekintés és a heurisztikus „guess" üzemmód is megengedett. Hasonlóan, a szegmentálás szavanként történik, és a legtöbb morfot eredményező első felbontást választjuk.

6.3.3. Morfessor – felügyelet nélküli statisztikai alapú morf szegmentálási módszerek

A Morfessor eszközök felügyelet nélküli adatvezérelt módszereket használnak, hogy a természetes nyelvek szóképzésének mögöttes szabályszerűségeit felfedjék. Vagyis nyelvi szabályok és adatbázis nélkül pusztán egy (gyakorisági vagy egyszerű) szótár alapján készítenek a szótárban szereplő szavak számára morf szegmentálást.

- ***MB (Morfessor Baseline)****:* [Creutz & Lagus 02], [Creutz & Lagus 05a]*:* A módszer célja olyan optimális morf lexikont és szegmentációt találni, amely tömör és hatékonyan írja le a bemeneti szótárat. A megközelítést a Minimum Description Length (MDL) alapelv [Risannen 78] inspirálta, amely szóköz nélküli mondatok szavakra szegmentálásánál [Brent 99], illetve hasonló lexikai modellezési feladatokban már jól teljesített [Goldsmith 01].

- ***MC-MAP (Morfessor Categories - Maximum A Posteriori)****:* [Creutz & Lagus 04], [Creutz & Lagus 05b]: a megközelítés célja a Morfessor Baseline által előállított szegmentáció finomítása. Az előző módszer kimenetének iteratív feldolgozása történik, miközben az egyes morf szegmentumok kategóriacímkéket kapnak, melyek a következők lehetnek: prefix (előtag), stem (tő), és suffix (toldalék). A címkék és soros függőségük felügyelet nélküli tanulása révén a szegmentáció pontosítható.

Noha mindkét eszköz képes a bemenő gyakorisági szótárnál a gyakorisági információt felhasználni, [Hirsimäki & Creutz+ 06] és saját tapasztalataink alapján ezt a lehetőséget nem érdemes használni. Így minden kísérletben uniform szóeloszlást adtunk meg bemenetként.

6.3.4. Kombinált statisztikai és szabály alapú morf szegmentáció

- ***CHM (Combined Hunmorph Morfessor):*** Ez a morf szegmentálási technika megpróbálja egyesíteni a szabály alapú és a statisztikai módszerek erényeit. A megközelítés lényege, hogy a Morfessor Baseline módszert annyiban módosítja, hogy az optimális morf szótár és szegmentáció keresésekor csak olyan morf szegmentációkat enged meg, amit a Hunmorph is elfogad. Vagyis, úgy is tekinthetjük, hogy a Hunmorph többes szegmentációinak dezambiguálását a Morfessor Baseline-ban implementált MDL alapelv szerint végezzük.

  Ezen felül, a megközelítés az önmagukban értelmes morf szegmentumokat akkor is engedélyezi, ha a Hunmorph ezeket az adott szó esetében nem fogadná el. További részletek a [C5]-ban találhatók az eljárásról.

E tanulmányban a CHM módszer felső referenciamódszerként használatos.

6.3.5. Morf alapú beszédfelismerés

A korábbi (1.2) definíciót szó helyett morf alapú beszédfelismeréshez ki kell terjeszteni. Hiszen nem elegendő a felismerésnél egy morf sorozat szövegszerű megjelenítése, mivel az írásunk szó alapú. Így morf lexikai egységek esetén az következő kiterjesztést alkalmazzuk. Használjunk explicit szóköz szimbólumokat (#) a szóhatárok visszaállításához a felismerési sztringben [Hirsimäki & Kurimo 04], és tekintsük ezeket a szimbólumokat is morfnak.
Ekkor a morfalapú beszédfelismerés a következőképpen formalizálható:

$$\hat{M} = \arg\max_{M} P(M)P(O \mid M) \tag{6.1}$$

$$\hat{W} = f(\hat{M}) \tag{6.2}$$

ahol *W* szósorozatot, *M* morf sorozatot, *O* akusztikus megfigyelés (jellemzővektor) sorozatot jelöl és *f* pedig egyszerű szövegösszefűzési és -törlési műveleteket a felismert morf sorozaton.

Ilyenkor az (1.5) helyett célszerű a következő akusztikus modell dekompozíciót használni:

$$\hat{M} = \arg\max_{M} P(M) \cdot P(\Phi \mid M) \cdot P(O \mid \Phi) \qquad (6.3)$$

A fonéma szinttől már ugyanúgy történhet az alacsony szintű akusztikai modellezés, mint korábban a szavaknál.

6.3.6. A morf szegmentációk alkalmazása

A morf szegmentáció célja tehát nem más, mint bármely *W* alapján *M*, vagyis f^{-1} meghatározása. Minden egyes morf szegmentációs eljárást az alábbiak szerint alkalmaztunk a magyar nyelvű, kézi erővel lejegyzett MALACH tanító szövegekre:

1. Az összes előfeldolgozott (kisbetűsített stb.) tanító szövegtokent egy listába gyűjtjük.
2. A betűzéseket, idegen szavakat, ill. nem szó tokeneket (pl. zaj jelöléseket) eltávolítjuk a listából.
3. A morf szegmentációt az adott módszerrel a szűrt listán végezzük el, vagyis ezen szavakat képezzük le morf sorozatokká. Az eredmény tehát egy $W \rightarrow M$ szótár, melyben csak bizonyos "reguláris-jellegű" szavak szerepelnek.
4. Az eredeti, szavakat tartalmazó tanítószövegbe beszúrjuk az explicit szóköz szimbólumokat (#).
5. A tanítószöveg azon szavait, melyek megtalálhatók a „reguláris" $W \rightarrow M$ szótárban a hozzájuk tartozó morf sorozatokkal helyettesítjük, a többi szót és # jelet meghagyjuk.

Ily módon a szó alapú tanítószöveget átalakítottuk egy morf sorozattá, ahol a „morf" megnevezést még szélesebb értelemben használjuk, mint korábban: a szavakat, morfokat, ragozott idegen szavakat és szóhatár-szimbólumokat mind egyenrangú morfként kezeljük. Ebből a morf-szintű tanító szövegből készítettük a felismerési szótárt és a nyelvi modelleket, amikor szó helyett morf alapú lexikai modelleket használtunk.

Fontos különbség a HUT (Helsinki University of Technology) módszerhez képest (pl. [Hirsimäki & Creutz+ 06]), hogy nem gyakoriság alapján távolítjuk el a morf szegmentáció tanulásához a nem illeszkedő szóalakokat, hanem a fenti módon. A [J1]-ben megmutatjuk, hogy a finn módszerhez képest jobb felismerési eredmények érhetők el az általunk javasolt technikával. Bár az annotáció során az idegen szavak megjelölése általában nem okoz jelentős plusz-munkát, automatikus módszerek alkalmazása is célravezető lehet ezek felderítésére, pl. [Szarvas & Farkas 06].

A különféle lexikai modellekkel elért beszédfelismerési eredmények az 6.2. táblázatban láthatók. A következőkben röviden összefoglaljuk az alkalmazott nyelvi és akusztikai modellezési megközelítéseket.

6.4. Nyelvi modellezés

Nyelvi modelltanítás céljára kizárólag a kézzel lejegyzett MALACH tanítószövegeket (mintegy 160.000 szövegtoken) használtuk. Mind a szó, mind a morf n-gram nyelvi modelleket a módosított, interpolált Kneser-Ney simítási technikával számítottuk az SRILM [Stolcke 02] nyelvi modellező eszköz segítségével. Az n-gram modellek fokszámát minden

lexikai modellezési megközelítésnél külön optimalizáltuk. Az SRILM eszköz alapértelmezett működésén felüli metszést (pruning) nem végeztünk a nyelvi modelleken.

6.4.1. Szó alapú nyelvi modellek

A szó alapú nyelvi modellt, *P(W)*-t, a legegyszerűbb „hagyományos" módon tanítottuk. Sem szó osztályokat sem címkéket nem használtunk, és természetesen explicit szóhatár szimbólumokat sem. Azon szóalakok száma, melyek a tanítószövegben csak egyszer fordultak elő az összes szóalakhoz képest (szótárméret) 62% volt, ami világosan mutatja a nyelvi modellezési nehézségeket agglutináló nyelv és viszonylag kisméretű tanítószöveg-adatbázis esetén. A felismerési hiba szempontjából az optimális n-gram fokszám a 3 volt.

6.4.2. Morf alapú nyelvi modellek

A morf nyelvi modelleket, *P(M)*-eket, az 6.3.6. szerint előkészített morf szövegadatbázissal tanítottuk. Meg kell említenünk, hogy az explicit szóköz szimbólumok bevezetése – mely a szóhatárok visszaállításához szükséges – rövidíti az n-gram modellek effektív előtörténetét (az *n*-1 hosszú „history"-t). Nyilvánvalóan akkor okozhat ez problémát, amikor a szavak átlagosan kevés morfra tagolódnak a morf szegmentálás során. Esetünkben az átlagos morf szám szavanként 1.6 körül mozgott mindegyik morf szegmentációs módszernél. A morf nyelvi modellek fokszáma 4 volt, mivel minden esetben ezekkel adódtak a legjobb felismerési eredmények.

6.5. Akusztikai modellezés

Mind a szó, mind a morf lexikai modellek esetén az akusztikai modellezés (*P(O|W)* és *P(O|M)* származtatása) két fő lépésből áll. Az első a lexikai egységek leképezése fonéma(szerű) sorozatokra, vagyis a kiejtésmodellezés. A kiejtés modellezése többféle problémát vet fel szavak és morfok esetén is – ezekkel a 7. fejezet foglalkozik tovább. A második lépés a fonéma alapegységek akusztikai modellezése, melynél nem tettünk különbséget a morf és szó lexikai modellek között.

6.5.1. Kiejtésmodellezés

Szó kiejtési modell. Azokat az ortografikus szavakat, melyek rendelkeztek egy vagy több fonológiai átirattal a lejegyzett tanítószövegben, kivételekként kezeltük. Ha csak egyféle fonológiai átirat szerepelt a tanítóanyagban, akkor ezt szerepeltettük a kiejtési szótárban. Ha többféle fonológiai átirat is szerepelt, akkor a tanítóadatok között mért előfordulási gyakoriságok szerinti ML súlyt kaptak az egyes alternatív ejtések. Bővebb információk a 7. fejezetben találhatók. A többi, nem kivételes ejtésmódú szót a [J2] szerinti eljárással, graféma-fonéma szabályok alapján képeztük le fonéma sorozatra.

Morf kiejtési modell. Mivel a tanítószöveg lejegyzése szó szinten történt, ezért a kivételes kiejtésű szavakból származó morfok kiejtésének származtatása nem triviális, gépi módszerekkel nem is oldható meg pontosan. Az újabb és újabb morf lexikai modellekre nyilvánvalóan nem praktikus új és új kivételszótárakat készíteni kézi erővel, ezért a kézi utófeldolgozást kizártuk a megoldási lehetőségek közül.

A kísérletekben a következő közelítő megoldást alkalmaztunk a morf szintű kiejtésmodellezésre. Azokat a morfokat, melyek a *szó* kivételszótárban előfordultak, a kivételszótárban szereplő kiejtési variánsokkal és súlyokkal modelleztük. A többit pedig ugyanazzal a szabály alapú eljárással képeztük fonéma sorozattá, amellyel a nem kivételes szavakat.

Ezzel a módszerrel számos morf kiejtési modellje elméletileg hibás vagy legalábbis szuboptimális lehet, azonban az eljárás gyakorlatilag további kézi munka nélkül lehetővé teszi a morf lexikai modellek alkalmazását a gépi beszédfelismerésben. A 7.1. táblázat megmutatja, hogy a szószintű kivételek jelentős részét sikerült így (újra)hasznosítani a morfok esetén is.

6.5.2. Fonéma alapú akusztikus modellek

Ugyanazokat a beszélőfüggetlen, környezetfüggő fonéma alapú akusztikus modelleket használtuk a szó és a morf lexikai modellek esetén. ML fonetikus döntési fákat használtunk a szóhatáron átívelő trifón modellek állapotainak csoportosítására. Így összesen mintegy 3000 HMM állapottal modelleztük alacsony akusztikai szinten a beszédhangokat. Három állapotú balról-jobbra HMM modelleket alkalmaztunk fonémánként, állapotonként maximum 10 Gauss-függvény szuperpozíciójával. Utóbbi alól kivétel a szünetmodell, amiben különböző zajmodelleket egyesítve 40 Gauss komponenst integráltunk. Az akusztikus modellek tanítása alapvetően a HTK-val történt [Young 06]. PLP (Perceptual Linear Prediction) lényegkiemelést használtunk akusztikai előfeldolgozás céljára, melyet egy módosított HTK eszközzel kaptunk [Psutka & Ircing+ 05]. Hangidőtartamokat nem modelleztünk explicit módon.

6.6. Kísérleti eredmények

6.6.1. Kísérleti beállítások

A felismerési tesztek során a dinamikus programozásra alapuló mintaillesztés a [C7]-ben bemutatott HMM-WFST alapú felismerő motor továbbfejlesztett változatával történt. A WFST felismerési hálózatot az AT&T FSM toolkitje segítségével építettük [Mohri & Pereira+ 02] a következő képlet szerint:

$$Felismerési\_hálózat = H \circ (C \circ det\,(L \circ G)) \qquad (6.4)$$

A általánosított trifón → fizikai HMM-állapot leképezés – melyet a *H* transzducer jelöl – a dekóderben történt. A fenti képlet érvényes úgy morf, mint szó lexikai modellek esetén, kis különbség mégis adódott a két hálózat között. Nevezetesen, míg a szó kiejtési modellek végéhez minden esetben hozzáillesztettünk egy opcionális szünet modellt, a morf kiejtési modellek esetén ettől eltekintettünk. Ehelyett a #, „szóköz” modellt magát képeztük le opcionális szünetmodellé. (Az opcionális szünetmodell funkcionálisan megfeleltethető a HTK [Young 06] szerinti „sp”, azaz „short pause” modellnek.)

A morf modellek esetében – azonos keresési mélység beállítások mellett – fennálló nagyobb keresési tér miatt számottevően lassabb a felismerés. Ezt a jelenséget azonban kompenzáltuk a keresési mélység csökkentésével és így nagy pontossággal azonos futási idő mellett vált lehetségessé a szó és morf modellek összehasonlítása. Az RTF=4.2-4.3 tartományon maradt minden felismerési tesztnél. A kísérleteket egyetlen 3GHz-es Pentium 4 gépen futattuk.

6.6.2. Beszédfelismerési eredmények

A beszélőfüggetlen felismerési tesztek során a teljes 8 órányi teszthalmazt felismertettük, és az eredményeket két részletben – illeszkedő és gyengén illeszkedő részhalmazra bontva –, valamint a teljes teszthalmazon értékeltük ki. A kiejtési és elemi akusztikus modellek tanítására a teljes (26 óra, 104 beszélő) tanító-adatbázist felhasználtuk.

Betűhibaarányokat (LER) is számoltunk, mert ezzel a mértékkel pontosabban lehet összehasonlítani az egyes a felismerési eredményeit, mint a hagyományos szóhibaarányokkal (WER).

A 6.2. táblázaton látható, hogy a legnehezebb a gyenge illeszkedésű részhalmaz felismerése. Itt egyetlen megközelítés sem tudott szignifikáns javulást elérni, még ha a javulás minden esetben pozitív is. Az illeszkedő részhalmazon viszont minden morf alapú felismerési eredmény szignifikánsan meghaladta a szó lexikai modellekkel elért eredményeket. A teljes teszthalmaz ad legmegbízhatóbb összehasonlítási alapot: betűhiba szempontjából minden morf alapú megközelítés szignifikánsan jobban teljesít, mint a szó alapú, ugyanakkor az általánosabban elfogadott WER tekintetében csak három megközelítésről mondható ez el.

6.2. Táblázat. Beszélőfüggetlen fonéma alapú beszédfelismerési eredmények a MALACH spontán magyar nyelvű beszédadatbázison különböző lexikai modellezési megközelítések mellett. A szó- és betűhibaarányok [%]-ban értendők, a referenciától szignifikánsan jobb értékeket dőlt betűkkel jeleztük.

Lexikai modellezési megközelítés	*Szótárméret*	*Gyenge illeszkedés*		*Illeszkedő részhalmaz*		Teljes teszt-halmaz	
		WER	*LER*	*WER*	*LER*	*WER*	*LER*
Szó – referencia	20k	56.1	28.6	52.9	25.5	54.5	27.1
MB (morf)	4.6k	56.0	28.5	*51.3*	*24.7*	53.6	*26.6*
MC-MAP (morf)	5.5k	55.8	28.2	*50.6*	*24.5*	*53.2*	*26.3*
HSF (morf)	8k	55.8	28.1	*51.3*	*24.5*	53.5	*26.3*
HCG (morf)	6.7k	55.4	28.0	*50.9*	*24.4*	*53.2*	*26.2*
CHM (morf)	6.7k	55.6	28.1	*50.5*	*24.1*	*53.0*	*26.1*

Fontos tehát kiemelni, hogy mind a statisztikai MC-MAP eljárás segítségével, mind a nyelvi tudással kialakított HCG, valamint a nyelvi és statisztikai módszert kombináló CHM eljárással kialakított morf alapú rendszer szóhibaarány tekintetében is szignifikánsan jobban teljesített, mint a szó alapú referencia. A megnevezett három megközelítés eredményei között nem volt szignifikáns különbség.

Fontosnak tartjuk megjegyezni, hogy noha a javulás mértéke szerény, az a tény, hogy morf lexikai modellekkel sikerült spontán nyelvű beszéd felismerését javítani (ráadásul szignifikánsan) nemcsak a magyar nyelv tekintetében, de nyelvtől függetlenül is újszerű eredménynek tekinthető. (Az első vonatkozó publikációink előtti időből [C3, B1] csak negatív eredményt sikerült fellelni spontán beszéd morf lexikai modellezésével kapcsolatban.)

6.6.3. Nemzetközi összehasonlítás

Esetlegesen felmerülhet a fenti eredmények értékelésénél, hogy a referenciául szolgáló szó alapú rendszer konstrukciójánál valamilyen hiba lépett fel, ezért alacsony a felismerési pontossága, és így annak megjavítása nem értékelhető megfelelően. Ennek az érvnek az ellentételezésére a 6.3. táblázatba foglaltuk össze a MALACH projekt más nyelveken, hasonló körülmények között elért legjobb (rész)eredményeit. A „baseline” jelző alatt a klasszikus felismerési megközelítést értjük a következő jellemzőkkel: ML tanítás, csak a lejegyzett adatbázis használatos tanításra, szó lexikai modellek.

6.3. táblázat. Nemzetközi „baseline” (viszonyítási alapként használt) felismerési eredmények és körülmények a MALACH projektben.

	Nyelv	DB méret [óra]	PP	OOV arány [%]	WER [%]
	Nagy adatbázissal				
Angol	[Byrne & Doermann+ 04]	200	86.9	8.2	53.3
Cseh	[Psutka & Ircing+ 05]	84	120	5.2	41.2
Szlovák	[Psutka & Ircing+ 05]	100	114	5.9	38.1
Orosz	[Psutka & Ircing+ 05]	100	122	5.1	46.8
	Kis adatbázissal				
Cseh	[Psutka & Ircing+ 02]	45	-	8.19	57.9
Orosz	[Psutka & Iljuchin + 03]	20	-	-	66.1
Magyar		26	644	14.5	54.5

Ahogy a 6.3 táblázat mutatja, a magyar nyelvnek a többinél nagyobb szóalaki változatosságából adódó nehézségeit tovább fokozza a lényegesen kisebb adatbázisméret. Látható, hogy a magyar nyelvű feladatnál a perplexitás kiugróan magas, továbbá a szótáron kívüli szavak aránya is csaknem kétszerese a második legnagyobb értéknek. Ezek a peremfeltételek különösen nagy kihívások elé állítják a beszédfelismerő rendszereket. Ennek ellenére a 200 órával tanított angol nyelvű eredményektől alig marad el a szóalapú magyar felismerési eredmény. Ha pedig a kisebb adatbázisokkal készült egyéb nyelvű eredményeket nézzük, a magyar egyértelműen jobban teljesít, mint akár az orosz, akár a jóval több adatot használó cseh rendszer. Az alapszintű eredmények tehát megerősítik azt a meggyőződésünket, hogy a kiindulási rendszer alaposan kidolgozott, mentes konstrukicós hibáktól, így annak szignifikáns megjavítása valós tudományos-technikai eredményként értékelhető.

Végül megjegyezzük, hogy az ebben az értekezésben nem tárgyalt, további beszélőadaptációs kísérletekben lényegesen nagyobb felismerési pontosságok mellett is sikerült a szó alapú megközelítéshez képest szignfikánsan jobb eredményeket elérni a morf lexikai modellezésnek köszönhetően [J1].

6.7. Összefoglalás

Bemutattuk eredményekeinket egy nemzetközi szinten ismert spontán, nagyszótáras, magyar nyelvű gépi beszédfelismerési feladaton. Megmutattuk, hogy szó helyett morf lexikai egységeket használva szignifikánsan javítható a beszédfelismerés hatásfoka. Felügyelet nélküli statisztikai módszerrel származtatott morf lexikai modellekkel is szignifikáns szó- és betűfelismerési hibacsökkenést sikerült elérni. A nyelvi adatbázison és szabályokon alapuló lexikai modellezési technikákhoz képest az adatvezérelt megközelítés nem teljesített szignifikánsan rosszabbul. Legjobb tudomásunk szerint az eredmények a magyar nyelvtől elvonatkoztatva is jelentős előrelépést jelentettek a spontán nagyszótáras folyamatos beszéd felismerésének területén.

7. Kiejtésmodellezés spontán magyar nyelvű beszéd gépi felismeréséhez

Mindeddig az ortografikus szó vagy morf lexikai alakok fonémákhoz hasonló egységekre bontását, vagyis a kiejtés (fonológiai szintű) modellezését nem vizsgáltuk meg közelebbről. Különösen spontán beszéd esetén azonban a lexikai alakok pontos, automatikus leképezése fonológiai egységekre korántsem triviális feladat, még „fonetikus" írásmódú nyelvek sem, mint amilyen a magyar.

A fejezetben alapvetően két kiejtésmodellezési megközelítést tárgyalunk, melyeket a MALACH spontán magyar nyelvű beszédadatbázison végzett beszédfelismerési tesztekkel értékelünk ki. Az egyik megközelítésnél graféma-fonéma szabályok alkalmazásával, illetve kivételes írás- vagy ejtésmódú (idegen, hagyományos stb.) szavaknál a kézi fonemikus átiratok felhasználásával törekszünk a kiejtési alternatívákat is megengedő leképezéseket készíteni fonéma alapegységekre. A másik megközelítésnél gyökeresen más megoldást alkalmazunk, fonémák helyett a lehető legegyszerűbb módon grafémákra azaz (egyjegyű) betűkre képezzük le nyelvi szabályok, kivételek vagy kézi átirat nélkül a lexikai egységeket. Végül olyan graféma akusztikus modellek alkalmazását vizsgáljuk magyar nyelvű spontán beszéd felismerésére, amely még a környezetfüggő modellezéshez szükséges döntési fák képzéséhez sem használ semmilyen nyelvi szabályt vagy nyelvi kategória-kialakítást, azaz a teljes akusztikus modellt tisztán adatvezérelt úton alakítjuk ki.

7.1. Bevezetés

Bár a magyar nyelvet fonetikai (vagy inkább fonológiai) írásmódúnak és a kiejtés modellezését graféma-fonéma szabályokkal jól kezelhetőnek tartjuk, néhány kérdés különösen a spontán beszéd felismerésével kapcsolatban felvetődik. Magyarban, hasonlóan más nyelvekhez, fontos szótagszintű törlések történhetnek a spontán beszéd során, melyeket a trifón modellezés már nem képes modellezni [Jurafsky & Ward+ 01]. Az idegen szavak jelentős részét, illetve az egyéb speciális kiejtésű szavak fonológiai szintű kiejtését sem tudjuk egyszerű szabályokkal modellezni. Az alternatív kiejtési változatok kezelése is erőforrásokat igényel. További nehézségeket jelenthet, ha – mint esetünkben is, a MALACH spontán magyar nyelű beszédadatbázisnál – a kivételes ejtésű szavakat ugyan ellátják fonetikus alakkal is, de ez szó szinten történik, míg a kiejtési modelleknek a morf lexikai egységekhez kell illeszkedniük.

Általában a beszédtechnológia szempontjából, de különösen a többnyelvű projektek szempontjából is kívánatos olyan kiejtésmodellezési technikákat megvizsgálni és alkalmazni, melyek a lehető legkevesebb kézi munkát igénylik, és a legkevésbé nyelvfüggőek. Ilyen például a graféma alapú kiejtésmodellezés [Kanthak & Ney 02], [Killer & Stüker+ 03], melyet tudomásunk szerint először mi alkalmaztunk mind magyar nyelvű folyamatos beszédfelismerésben, mind a nemzetközi MALACH projektben. (Magyar nyelvű graféma alapú izolált szavas felismerésre van korábbi példa [Zgank & Kacic+ 05], azonban csupán kezdeti, referencia nélküli eredményeket adnak meg.)

A következőkben a magyar nyelvre optimalizált graféma-fonéma szabályokat alkalmazó fonéma alapú kiejtésmodellezést hasonlítjuk össze a nyelvi szabályoktól nagyrészt, illetve teljesen mentes graféma alapú kiejtésmodellezési megközelítésekkel a spontán magyar MALACH adatbázison végzett beszédfelismerési kísérletekben.

7.2. Az automatikus fonológiai kiejtésmodell-előállítás problémái

Az angol és az egyéb, nem fonetikus írásmódú nyelvek esetén bevett gyakorlat, hogy az egyes szavakhoz kézi erővel készítik a különböző fonológiai átiratokat, majd a tanító-adatbázison a kényszerített felismerés módszerével választják ki az aktuális fonológiai realizációkat [Young 06]. E megközelítés hátránya nyilvánvaló, a kiejtési szótáron kívüli elemeket nem lehet automatikusan átírni, valamint akkor sem használható jól, ha az adott szóhoz nem tartozik hullámforma. Ezért számos „intelligens", illetve statisztikai módszert dolgoztak ki, melyek segítségével például a már rendelkezésre álló átiratokból tanulva új szavak fonemikus átirata becsülhető, pl. [Torkkola 93], [Besling 94], [Suontasuta & Hakkinen 00].

Magyar nyelv esetén nem tűnik szükségesnek ilyen módszereket alkalmazni, hiszen hasonlóan a finnhez vagy észthez [Kurimo & Puurula+ 06] a fonetikus írásmód miatt a magyar szavak túlnyomó többségénél legalább a „kanonikus" kiejtés könnyen előállítható automatikusan is. A gyakorlatban azonban – a látszólagos könnyűség ellenére – a szó vagy morf lexikai egységek automatikus fonológiai átírása számos problémát vet fel [B4], [Zsigri & Tóth+ 04], melyeket a következőkben részletezünk.

7.2.1. Kiejtési kivételkezelés

Nyilvánvaló, hogy a kivételes kiejtésű szavakat kivételszótárral kell kezelni, amelynek összeállítása történhet akár kézzel is. A problémát az jelenti, hogy hogyan detektáljuk, hogy kivételes szóról, vagy szórészletről van szó.

Tekintsük például az alábbi kivételszótár-bejegyzést:

Weimar → v e j m á r
Ostrava → o sz t r a v a
gipsy → dzs i p sz i

Hogyan alkalmazzuk ezeket a következő szóalakok esetén:

weimarizálódás, Ostravába, legipsyzte?

A kisbetűs rész-sztringre illeszkedés láthatóan nem jelent általános megoldást.

7.2.2. Graféma-fonéma konverzió

Még ha csak a valóban fonológiai írásmódú szavak kanonikus kiejtésmodellezésére szorítkozunk, akkor is alapvető problémák léphetnek fel a többjegyű betűk miatt.

Az első részprobléma a *tokenizáció*, vagyis a konvertálandó grafémacsoportok azonosítása.

Pl. ha a graféma-fonéma szabályok az alábbiak:

s → s
zs → zs
c → c
cs → cs
ggy → ggy
gy → gy
g → g
stb.

a

„lecsós malacsült”,
„rozsdás községtábla”,
„meggyel meggyógyít”

szövegrészletek további nyelvi, akusztikai információk nélkül nem képezhetők le helyesen a nyers fonológiai szintre. Nyers fonológiai szint alatt a *beszédhangszándéknak* megfelelő mögöttes fonológiai formát, és nem a felszíni, fonológiai koartikulációkat tartalmazó fonémasorozatot értjük.

7.2.3. Morf lexikai modellek

A morf lexikai modellek használata néhány problémát megold, azonban nem mindet, és újak is keletkeznek.

Pl. a „község” a morfológiai analízisek során tipikusan nem bomlik „köz + ség” alakra, ilyenkor továbbra sem oldható meg kivételszótár nélkül a helyes fonéma sorozatra való leképzése.

A gyakorlatban azonban nagyobb problémát jelenthet a fordított eset, amikor is a morfológiai szegmentáció – akár statisztikai alapú, akár szabály alapú – a többjegyű betűket felbonthatja, és így a helyes graféma-szabályok sem érvényesülhetnek.

Pl. válasszon → válas + szon, arannyal → aran + nyal, meggyel → meg + gyel

7.2.4. Fonológiai koartikulációk

Korábban a beszédfelismerésben nagy jelentőséget tulajdonítottak a fonológiai koartikulációk jelölésének a fonológiai átiratban. Ez azonban csak szóbelsőben volt megoldható szótárakban formalizálva. Amint az 5. fejezetben megmutattuk, még a szóhatáron átívelő fonológiai koartikulációs jelenségek explicit modellezése sem hozott javulást az alacsonyabb szinten történő (állapotcsoportosított trifón) modellezéshez képest. Így a fonológiai koartikuláció explicit jelölését beszédfelismerés esetén szükségtelennek, sőt károsnak tartjuk és nem is alkalmazzuk. Ugyanis a nyers fonológiai alakokból az 5.3-ban részletezett P transzducerrel a fonológiai koartikulációk jórészét tartalmazó fonémahálózat könnyen előállítható, fordítva viszont ez általában nem lehetséges.

Pl. értsd → é r dzs d, értsd te is → é r cs t e i s

7.2.5. Spontán beszédre jellemző kiejtési variációk előállítása

Mint ismeretes, a spontán beszédben a leírt alaknak megfelelő fonológiai realizációhoz képest jelentős változások történhetnek, melyek a standard fonológiai koartikulációkkal nem írhatók le. Itt elsősorban hangkiesésekre gondolunk, de nem ritka a teljes szótag kiesése, sőt akár több szótag is törlődhet, illetve az artikuláció is jelentősen torzulhat. Ezen jelenségek egy része a fonológiai szinten megragadható, azonban az automatikus modellezésük, algoritmizált előállításuk nem tűnik egyszerűen megoldhatónak. Így inkább a kézi átiratból származó információk felhasználásának módja vetül fel megoldandó problémaként.

Pl. azt mondja → a sz o n gy a, miért → m é r t, m é, majd → m a j

7.3. Kiejtésmodellezési megközelítések

Két közelítő jellegű megoldási lehetőséget javaslunk az előzőekben említett problémák kezelésére. Fonéma alapú kiejtésmodellezésnél igyekszünk – pótlólagos kézi munka nélkül – megfelelni a nyelvészeti elvárásoknak. Graféma alapú modellezésnél kiejtési kivételkezelés helyett az elemi akusztikus modelleket közvetlenül betűkre építve, tisztán mérnöki-statisztikai megközelítéssel vágjuk át a gépi beszédfelismerés „gordiuszi csomóját".

7.3.1. Fonéma alapú kiejtésmodellezés

Ennél a megközelítésnél a szó vagy morf lexikai modellek leképezése fonéma (vagy inkább fonémaszerű) fonológiai egységekre történik.[13]

A feladat tehát adott szó vagy morf sorozat leképzése fonémasorozatra, illetve alternatív, valószínűségekkel ellátott fonémasorozatokra, vagyis a $P(\Phi|W)$ ill. $P(\Phi|M)$ meghatározása.[14] A leképezést lexikai egységekként végezzük, a szó/morf környezettől való fonológiai szintű kiejtési függőséget nem modellezzük.

A MALACH adatbázis esetén az alábbi kétszintű, fonéma alapú kiejtésmodellezést alkalmaztuk.

1. Elsődleges kiejtési kivételkezelés

Azon szavak, melyek a kézi lejegyzés során egy vagy több fonológiai alakkal is el lettel látva egy elsődleges kivételszótárba kerültek. Az egyes kiejtési variánsokhoz tartozó valószínűségeket az annotáció alapján számított relatív gyakoriságokkal becsültük.

Illusztráció:	w	$P(\Phi^i \mid w)$	Φ^i
	miért	0.011	m é
	miért	0.426	m é r
	miért	0.269	m i é r
	miért	0.292	m i é r t

Az elsődleges kivételszótárat mind szó, mind morf lexikai egységek esetén ugyanúgy alkalmaztuk. Ha egy szótárelem ortografikus alakja pontosan egyezett a kivételszótár baloldalán álló alakkal, akkor a fonológiai kiejtési modellje pontosan a szótárban megadott, valószínűségekkel ellátott kiejtési változatokat tartalmazó forma lett.

2. Graféma-fonéma szabályok alkalmazása

Az általános graféma-fonéma szabályok a magyar nyelvi fonológiai írásmódnak többé-kevésbe megfelelő szavak kiejtésének egyértelmű – kiejtési variánsok nélküli – leképezésére szolgálnak. Ugyanakkor, a másodlagos kiejtési kivételkezelés is ezek által a dinamikusan bővíthető graféma-fonéma szabályok által valósul meg.

Egy adott, az elsődleges kivételszótárban nem előforduló szó vagy morf lexikai egység leképezésére a következő eljárást alkalmazzuk:

[13] Az általunk használt rendszerekben a hosszú és rövid mássalhangzók megkülönböztetése nem történik meg, ennek ellenére, az egyszerűség kedvéért, mint fonéma egységekre fogunk hivatkozni rájuk.
[14] V.ö.: (1.2), (1.5) és (6.3).

- Az általános kiejtésmódú szavakat és a másodlagos kiejtési kivételeket ugyanúgy a graféma-fonéma szabályokkal kezeljük.

  Pl. graféma sorozat → fonéma sorozat

  #Ostrav → o sz t r a v
  zs → zs
  község → k ö z s é g
  th# → t
  x → k sz
  #x# → i k sz
  w → v

  (# a szóhatár szimbólumot jelöli, az ABC minden betűjét leképezzük valamilyen fonémára ill. fonémasorozatra)

- Az adott bemenő ortografikus lexikai egység elejére megkeressük az első leghosszabban illeszkedő graféma-fonéma szabály baloldalt. Ezt lecsatoljuk, a fonémasorozatot eltároljuk és az eljárás folytatódik tovább, amíg a bemenő sztring el nem fogy.

  Illusztráció:

  Ostravából → o sz t r a v á b ó l
  kisközségbe → k i s k ö z s é g b e
  wolf → v o l f
  Tóth → t ó t

Mint látható, viszonylag egyszerű kiejtési kivételkezeléssel és graféma-fonéma szabályrendszerrel is hatékonyan kezelhető a 7.2.-beli problémák jó része. A gyakorlatban azonban mindig találhatunk olyan példát, amire sem a kivételszótárak, sem a szabályrendszer nincs felkészülve, és így az előállított kiejtés sem lesz helyes. Ez velejárója a „szakértői” szabály alapú rendszereknek.

7.3.2. Graféma alapú kiejtésmodellezés

Mint említettük, egyre jobban terjednek a részben vagy teljesen statisztikai alapon működő graféma-fonéma átalakító rendszerek. Ilyen például a [Suontasuta & Hakkinen 00], ahol döntési fa alapú graféma-fonéma átalakítást végeznek. [Kanthak & Ney 02] felteszi a kérdést, hogy ha a környezetfüggő fonéma modellek állapotcsoportosítására is döntési fákat használunk,[15] miért ne lehetne összevonni a két lépést, és a döntési fákkal közvetlenül a környezetfüggő grafémákat akusztikusan modellezni. A beszédfelismerési kísérletek pozitív választ adtak a felvetésre. A fonetikus kérdésekből származtatott graféma döntési fákkal [Kanthak & Ney 02] lényegében ugyanolyan jó felismerési pontosságokat lehetett elérni német, holland, és olasz nyelvre, mint a hagyományos környezetfüggő fonéma alapú megközelítéssel. [Killer & Stüker+ 03] még tovább megy, megmutatták, hogy német és spanyol nyelvre a minimalista szingleton graféma osztályokkal – amikor grafémikus *környezet gyanánt* minden osztály egy elemet tartalmaz csupán, graféma típusonként egyetlen (egyjegyű) betűt – szintén lényegében ugyanolyan jó eredmények érhetők el, mint kifinomult

[15] Lásd a 4. fejezet vonatkozó részeit.

kézi vagy automatikusan generált kérdésekkel. Ezáltal a kiejtésmodellezés teljes mértékben nélkülözni tudja a nyelvfüggő szabályokat.

A (környezetfüggő) graféma alapú kiejtésmodellezés ugyanakkor nem teljesített jól angol és francia nyelv esetén. Vagyis, minél inkább jellemző a fonetikus írásmód egy nyelvre, annál inkább várható a megközelítés sikere. Így természetesen adódott, hogy magyar nyelvre is megvizsgáljuk az alkalmazását.

A graféma alapú akusztikai és kiejtésmodellezésnek a MALACH magyar nyelvű adatbázis esetén az ad különös jelentőséget, hogy mint az 5. fejezetben láttuk, a morf alapú lexikai megközelítések egyértelműen jobban teljesítettek, mint a szó alapúak. A morf szegmentálások azonban – főként a tisztán statisztikai, de a szabály alapú is – sok esetben egyetlen fonémát kódoló grafémasorozatot is felbontanak, így fonológiai szempontból hibás szegmentációt eredményeznek. Graféma alapú akusztikus és kiejtési modellezésnél ilyen probléma nem merül fel. Továbbá, az idegen szavak ejtésmódját is megtanulhatja a graféma akusztikus modell, így az ezekhez kapcsolódó kivételszótárra sincs szükség.

A graféma akusztikai modellek lényegi jellemzője a triviális kiejtési modell előállítás, ezért nem illeszkedik a megközelítéshez a spontán ejtésmód esetén a többes kiejtési variációk modellezése. A kísérletekben ezért egyáltalán nem alkalmaztunk kézi fonológiai alakokat, csak és kizárólag ortografikus lexikai alakokból, kivétel szótárak és nyelvi szabályok nélkül származtattuk a graféma alapú kiejtésmodellezést.

A megközelítés *formális definíciója* morf lexikai modellek esetén tehát a következő:

$$\hat{M} = \arg\max_{M} P(M)B(\Gamma \mid M)P(O \mid \Gamma) \tag{7.1}$$

ahol Γ graféma (egyjegyű betű) sorozatot jelöl, a *B* pedig bináris (0 vagy 1 kimenetű) valószínűségeloszlást. Ezt a megközelítést „morf-graféma" alapú beszédfelismerésnek is hívjuk a továbbiakban [C3]. [16]

A *B(Γ|M)* graféma kiejtési modell tehát az ortografikus alak betűinek triviális sorozatára képez. Illusztráció:

Tóth → t ó t h
Churchill → c h u r c h i l l
tavaly → t a v a l y
aran | nyal → a r a n n y a l

Az elemi graféma akusztikai modellek a graféma környezetükből „tanulják meg", hogy milyen akusztikai eloszlásokat kell alkalmazni.

[16] A szó-graféma megközelítés esetén a képlet teljesen hasonló – csupán az *M*-eket kell *W*-re cserélni – ezt külön nem írjuk fel.

7.4. Kísérleti eredmények

A következő beszédfelismerési kísérletekben a graféma alapú modellekkel elért eredményeket hasonlítjuk össze a korábban már közölt fonéma alapú eredményekkel.

7.4.1. Adatbázis-jellemzők és kísérleti beállítások

A gépi beszédfelismerési kísérleteket ugyanazzal a MALACH spontán magyar nyelvű adatbázissal, és mindenben hasonlóan végeztük, mint az előző fejezetben említetteket. Ezért az adatbázis és általános paraméterek újbóli bemutatását mellőzzük, csak a jelen fejezet szempontjából fontos adatokat részletezzük, illetve a korábbi tanítás-tesztelési felállástól való eltérést ismertetjük.

7.1. Táblázat. A magyar MALACH adatbázis esetén a szakértői kézi címkézés alapján számolt kivétel- és súlyozott kivételszótárak mérete és fedése a tanító-adatbázison. A kivételszótár részét képezi a súlyozott kivételszótár.

Lexikai modell típus	Teljes szótár mérete	Kivételszótár		Súlyozott kivételek szótára	
		Méret	Fedés [%]	Méret	Fedés [%]
Szó	20k	1743	47.1	720	46.2
MB (morf)	4.6k	521	32.9	199	32.6
MC-MAP (morf)	5.5k	492	27.3	163	26.9
HSF (morf)	8k	576	25.6	243	25.3
HCG (morf)	6.7k	539	23.1	216	22.9
CHM (morf)	6.7k	565	29.6	224	29.3

Ahogy a 7.1. táblázat mutatja, a 7.3.1-ben ismertetett kiejtési kivételkezelési technikával a szó alapú kivételek jelentős részét sikerült újrahasznosítani morf lexikai modellek esetén is. Az is jól látható, hogy a kivételes ejtésmódú szavakat – és ezeken belül is a kiejtési alternatívákat is igénylő kifejezéseket – gyakran használták az adatközlők. Tehát (fonológiai) kiejtésmodellezés szempontjából semmiképpen sem nevezhető könnyűnek a beszédfelismerési feladat.

A felismerési paraméterek, használt eszközök megegyeztek a korábbiakkal (3-gram szó nyelvi modell, 4-gram morf nyelvi modell, 3 állapotú HMM fonéma modell, 10 komponensű GMM állapotonként, ML döntési fa alapú trifón állapotcsoportosítás stb.). A graféma akusztikus modellek ugyanannyi, 3000 körüli HMM állapotot tartalmaztak, mint a fonéma alapúak. Kisebb különbség volt, hogy az annotációban a zajjelöléseket teljesen figyelmen kívül hagytuk, és a szünet- vagy „szóköz” modell így 10 Gauss függvényből állt csak. A környezetfüggő graféma modellekhez a döntési fa építésénél a fonéma akusztikus modelleknéhez nagyon hasonló, abból triviálisan származtatott fonológiai-fonetikai kategóriákat használtunk (egy adott fonéma osztály fonémáihoz használt grafémák alkották a megfelelő graféma osztályt [Kanthak & Ney 02]). A szingleton graféma modellekhez pedig nem használtunk semmilyen nyelvspecifikus tudást, azaz a döntési fa építésénél az egyelemű (szingleton) graféma osztályok csupán az egyjegyű betűk felsorolását jelentették és adták a grafémikus környezetet.

Megemlítendő, hogy adott keresési mélység beállítások mellett a graféma alapú felismerő-rendszerek általában gyorsabbak voltak, mint a fonéma alapúak. Ennek elsődleges oka a kiejtési változatok hiánya miatti kisebb keresési tér. A jelenséget azonban mind szó, mind morf lexikai modellek esetén kompenzáltuk a keresési mélység megfelelő növelésével, így az RTF itt is 4.2 és 4.3 közöttire korlátozódott. Tehát bármely két, a MALACH adatbázison bemutatott felismerési eredmény egzakt módon egymáshoz hasonlítható.

7.4.2. Graféma alapú beszédfelismerési eredmények

A 7.2. táblázaton mutatjuk be az új beszédfelismerési eredményeket. A jobb szélen található „relatív javulások" oszlopban az egyes lexikai modellezek fonéma alapú megfelelőjéhez viszonyítottunk a teljes tesztadatbázis eredményeit figyelembe véve.

7.2. Táblázat. Beszélőfüggetlen graféma alapú beszédfelismerési eredmények a MALACH spontán magyar nyelvű beszédadatbázison különböző lexikai modellezési megközelítések mellett. A *szó-fonéma* referenciához viszonyított szignifikáns javulásokat dőlt betűvel jelöltük. A szó-, betűhibaarányok és a teljes teszthalmazon a *fonéma alapú, ugyanazon lexikai modellhez viszonyított relatív javulások* [%]-ban értendők.

Modellezési megközelítés	*Gyenge illeszkedés*		*Illeszkedő részhalmaz*		Teljes teszt-halmaz		Relatív javulás	
	WER	*LER*	*WER*	*WER*	WER	LER	WER	LER
Szó – graféma	57.0	28.6	53.8	25.7	55.4	27.1	*-1.6*	0
MB – graféma	57.2	28.5	51.8	*24.8*	54.4	26.6	-1.4	0
MC-MAP – graféma	56.1	28.2	51.5	*24.8*	53.8	*26.5*	-1.1	-0.7
HSF – graféma	56.6	28.1	51.6	*24.6*	54.1	*26.3*	-1.1	0
HCG – graféma	56.5	28.1	51.4	*24.5*	53.9	*26.3*	-1.3	-0.4
CHM – graféma	56.4	28.2	*51.0*	*24.3*	53.7	*26.2*	-1.3	-0.4
MC-MAP – graféma sz.	55.9	*27.9*	51.5	*24.8*	53.7	*26.4*	-0.9	-0.3

Látható, hogy a graféma alapú modellezés hatására egyetlen morf lexikai modellezési megközelítésnél sem nőtt szignifikánsan a teljes teszthalmazon mért beszédfelismerési hiba. A szó lexikai modellnél ugyan szignifikánsan romlott a szófelismerés pontossága, ugyanakkor a betűfelismerési pontosságot a graféma alapú akusztikus modellezés itt sem rontotta érzékelhetően.

A *szó-fonéma referenciához* hasonlítva az eredményeket látható, hogy számos részeredmény, illetve a teljes teszthalmazon mért betűhibaarányok jelentős része még javult is dacára a kivételszótárak, graféma-fonéma átalakítási szabályok teljes kiiktatásának.

A Morfessor Categories-MAP lexikai modellel és graféma szingleton akusztikus modellel készült teljesen adatvezérelt megközelítésről elmondható, hogy graféma modellek mellett teljesítménye alig marad az ilyenkor is legjobban teljesítő CHM technikától. Mi több, a teljesen adatvezérelt megközelítés a legnehezebb felismerési feladatnak számító „gyenge illeszkedésű" teszthalmazon a 27.9%-os betűhibaaránnyal abszolút csúcsot állított fel.

7.4.3. Következtetések

Az eredmények alapján levonhatjuk a következtetést, hogy létre lehet hozni *teljesen adatvezérelt módon* is a klasszikus szó-fonéma alapú beszédfelismerési megközelítéssel versenyképes beszédfelismerő rendszert. Ahogy a 7.3. táblázat szemlélteti, a felügyelet nélküli statisztikai eljárással kialakított MC-MAP morf lexikai modellek a szingleton graféma modellekkel együtt minden részhalmazon jobban teljesítenek mint a szó-fonéma alapú referencia, sőt a betűhibaarányok tekintetében szignifikáns is a javulás.

7.3. Táblázat. A klasszikus (szó-fonéma) és a teljesen adatvezérelt (MC-MAP – graféma szingleton) beszédfelismerési megközelítések eredményeinek összevetése a MALACH spontán magyar nyelvű beszédadatbázison. A szignifikáns javulásokat dőlt betűvel jelöltük, a szó-, betűhibaarányok és relatív javulások [%]-ban értendők.

Modellezési megközelítés	*Gyenge illeszkedés*		*Illeszkedő részhalmaz*		Teljes teszt-halmaz	
	WER	*LER*	*WER*	*WER*	WER	LER
Szó – fonéma (referencia)	56.1	28.6	52.9	25.5	54.5	27.1
MC-MAP – graféma szingl.	55.9	27.9	51.5	24.8	53.7	26.4
Relatív javulás	0.4	*2.5*	2.7	*2.8*	1.5	*2.6*

Alább röviden szemléltetjük, hogy a szó-fonéma alapú referencia rendszer milyen jellegű nyelvspecifikus szabályokat alkalmaz, azaz hogy mire nincs szükség versenyképes beszédfelismerési eredményekhez:

- Szó-fonéma alapú megközelítés (referencia):
  - Valószínűségi súlyozású alternatív kiejtések, pl.
    miért 0.011 m é
    miért 0.426 m é r
    miért 0.269 m i é r
    miért 0.292 m i é r t
  - Idegen és hagyományos írású szavak kivételes kiejtései, pl.
    Churchill cs ö r cs i l
    Kossuth k o s ú t
  - Graféma-fonéma átalakítási szabályok, pl.
    cz c
    ch# cs
    ck# k
    ly j
    (# a szóhatár szimbólumot jelöli)
  - Fonetikai kategóriák, pl.
    NASAL: m, n, ny
    FRONT: e, é, i, í, ö, ő, ü, ű

Ezekkel szemben az MC-MAP – graféma szingleton alapú megközelítés semmilyen fenti, sem egyéb nyelvspecifikus szabályt nem használ, hiszen nemcsak a nyelvi és lexikai modell, de az akusztikus modell is teljesen adatvezérelten készül, melyek képzéséhez mindösszesen csak a tanítófelvételek hullámformáira és azok (ortografikus) szöveges lejegyzéseire van bemenetként szükség. A következtetések beszélőadaptáció mellett is érvényesnek bizonyultak [J1].

7.5. Összefoglalás

Különböző kiejtési-, akusztikai modellezési megközelítéseket mutattunk be spontán, nagyszótáras magyar nyelvű beszédfelismerési alkalmazásban. Először a fonéma alapú kiejtésmodellezést tárgyaltuk, melynél kézi erővel előállított kiejtési variánsokat, illetve kivételszótárat, valamint graféma-fonéma szabályokat alkalmaztunk a lexikai egységek fonémasorozatra való leképezésére. Majd bevezettük a graféma alapú kiejtési és akusztikai modellezést, ahol minden előbb említett szabálytól és erőforrástól megválva gyakorlatilag ugyanolyan jó eredményeket értünk el, mint a fonéma alapú rendszereknél. Végül a graféma akusztikus modellek képzésénél használt nyelvi ismereteket igénylő fonetikai kategóriákat is kiküszöböltük. Végeredményben a teljesen adatvezérelt – de a magyar nyelv morfológiai jellegét a statisztikai morf modellek révén figyelembe vevő – technikával minden tekintetben jobb, a betűhibaarányokat tekintve szignifikánsan jobb beszédfelismerési eredményeket értünk el a számos nyelvi szakértői tudást alkalmazó szó-fonéma alapú referenciánál.

8. Összefoglalás, tézisek

Értekezésemben bemutattam főbb tudományos eredményeimet a magyar nyelvű gépi beszédfelismerés területén. Összességében arra jutottam, hogy nyelvspecifikus szakértői szabályok[17] alkalmazása nem nélkülözhetetlen a gépi beszédfelismeréshez. Ellenkezőleg, megmutattam, hogy teljesen adatvezérelten – azaz mélyebb nyelvi szakértelem nélkül is – kompetitív beszédfelismerő rendszer készíthető az adott, spontán, magyar nyelvű feladatra. Mindez nem jelenti azt, hogy a magyar nyelv sajátosságaira (elsősorban a toldalékoló jellegre és a döntően fonológiai írásmódra) felesleges lenne tekintettel lenni. Értékelésem szerint inkább arról van szó, hogy a magyar nyelv globális jellemzőit figyelembe véve a megfelelő adatvezérelt technikákat választhatjuk ki, amelyekkel már ténylegesen versenyképes beszédfelismerési alkalmazások fejleszthetők ki, relatíve gyorsan és kisebb erőforrásokkal. A következőkben téziscsoportonként összegzem tudományos állításaimat, következtetéseimet, majd röviden kitérek az eredmények további alkalmazásaira is.

I. téziscsoport: A fonetikai koartikuláció modellezése magyar nyelvű beszéd gépi felismeréséhez

I.1. tézis: [B2, B3, C7, C8] *Kísérleti úton megmutattam, hogy visszametszéses trifón állapotcsoportosítású környezetfüggő beszédhangmodellekkel elérhető szignifikáns beszédfelismerési pontosságjavulás magyar nyelven a környezetfüggetlen beszédhangmodellezéssel elért eredményekhez képest.*

I.2. tézis: [B2]. *Kísérleti úton megmutattam, hogy az alapvetően nyelvi szabályok által vezérelt visszametszéses trifón állapotcsoportosítású beszédhangmodellekkel elért beszédfelismerési pontosságoknál elérhető szignifikánsan jobb eredmény a jelentősebb mértékben statisztikai elvű, ún. maximum likelihood fonetikus döntési fa alapú trifón állapotcsoportosítási módszer [Young & Odell+ 94] alkalmazásával.*

II. téziscsoport: A fonológiai koartikuláció (hasonulási jelenségek) modellezése magyar nyelvű beszéd gépi felismeréséhez

II.1. Tézis: [J2, B3, B4, C6, C9, C10] *Kísérleti úton megmutattam, hogy – amennyiben az akusztikus modellek tanításakor a tanító-adatbázisban a fonológiai koartikuláció figyelembe lett véve (például kézi fonetikus átirat révén) – a felismerési tesztekben egyes tipikus fonológiai koartikulációs jelenségek explicit (szóhatárokon is átívelő) modellezésével elérhető szignifikánsan magasabb beszédfelismerési pontosság, mint a jelenség tekintetbe vétele nélkül.*

II.2. Tézis: [C6], *Kísérleti úton megmutattam, hogy implicit fonológiai koartikuláció-modellezéssel – amikor is mind az akusztikus modellek tanításakor, mind a felismerési tesztek során eltekintünk a fonológiai koartikuláció jelenségétől – elérhető kompetitív (nem szignifikánsan alacsonyabb) beszédfelismerési pontosság ahhoz képest, mint amikor a fonológiai koartikulációs jelenségek jelentős részét expliciten modellezzük mind tanítás, mind tesztelés során.*

[17] Ilyenek például a nyelvtani, morfológiai (morfo-szintaktikai) szabályok, a fonológiai koartikulációs szabályok, a kiejtési kivételszabályok vagy éppen az egyes beszédhangok fonetikai kategorizálásank szabályai.

Következmény: *A tipikus fonológiai koartikulációs jelenségek explicit modellezése nem nélkülözhetetlen a magyar nyelvű gépi beszédfelismerésben, hiszen az explicit modellek körülményes integrációja elhagyható anélkül, hogy a felismerési pontosság feltétlenül szignifikánsan csökkenne.*

III. téziscsoport: Spontán magyar nyelvű beszéd lexikai modellezése gépi beszédfelismeréshez

III.1. tézis: [J1, B1, C1, C2, C3, C4, C5] *Kísérleti úton megmutattam, hogy spontán beszéd gépi felismerése esetén szó helyett kisebb, morfémaszerű (továbbiakban: morf) lexikai egységek megfelelő alkalmazásával elérhetők szignifikánsan magasabb felismerési pontosságok.*

III.2. tézis: [J1, B1, C1, C2, C5] *Kísérleti úton megmutattam, hogy spontán beszéd gépi felismerése esetén felügyelet nélküli statisztikai módszerrel [Creutz & Lagus 05b] származtatott morf lexikai egységek alkalmazásával elérhetők a szó alapú megközelítés eredményeitől szignifikánsan magasabb, a nyelvi szabály alapú ill. kombinált (statisztika + nyelvi szabályok alapján előállított) morf megközelítések eredményeitől pedig nem szignifikánsan alacsonyabb felismerési pontosságok.*

IV. téziscsoport: Spontán magyar nyelvű beszéd akusztikai és kiejtésmodellezése gépi beszédfelismeréshez

IV. 1. tézis: [J1, C3] *Kísérleti úton megmutattam, hogy spontán magyar nyelvű beszéd gépi felismerése esetén környezetfüggő graféma (alfabetikus karakter) alapú akusztikus modellezéssel elérhető nem szignifikánsan alacsonyabb felismerési pontosság, mint fonéma alapúval (morf lexikai modellezés mellett).*

Következmény: *Kézi kivételszótárak és graféma-fonéma átalakítási szabályok alkalmazásának hiánya nem feltétlenül okoz szignifikáns felismerési pontosságromlást magyar nyelvű gépi beszédfelismerésnél. Ezek a nyelvi tudásforrások tehát nem tekintendők nélkülözhetetlennek a magyar nyelvű gépi beszédfelismerésben.*

IV. 2. tézis: [J1] *Kísérleti úton megmutattam, hogy spontán magyar nyelvű beszéd gépi felismerése esetén környezetfüggő ún. graféma-szingleton alapú akusztikus modellezéssel – amikor is az alkalmazott ML döntési fa alapú trifón állapotcsoportosításnál csupán triviális, egyelemű graféma osztályokat definiálunk környezet gyanánt – elérhető nem szignifikánsan alacsonyabb felismerési pontosság, mint fonéma alapú akusztikus modellekkel (morf lexikai modellezés mellett).*

Következmény: *Nyelvspecifikus szabályok és szakértői nyelvi tudás explicit alkalmazásának hiánya nem feltétlenül okoz szignifikáns pontosságcsökkenést magyar nyelvű gépi beszédfelismerés esetén. A nyelvi szakértői tudás és a nyelvspecifikus szabályok explicit alakjukban (fonetikai osztálydefinícók, kiejtési és betű-hang átalakítási szabályok, szótárak stb.) nem tekintendők tehát a magyar nyelvű gépi beszédfelismerés létfontosságú kellékeinek.*

V. téziscsoport: Spontán magyar nyevű beszéd felismerése explicit nyelvi szabályok nélkül (szintézis)

V. 1. tézis: [J1] *Kísérleti úton megmutattam, hogy spontán magyar nyelvű beszéd gépi felismerésénél explicit nyelvi ismeretek alkalmazása nélkül*[18] *is elérhető kompetitív felismerési pontosság a klasszikus szó-fonéma alapú megközelítéshez képest, mely számos nyelvspecifikus szakértői tudás*[19] *alkalmazását igényli.*

A tézisekben összefoglalt eredmények egyrészt a magyar nyelvű beszédfelismerési fejlesztéseket tehetik hatékonyabbá (gyorsabbá és olcsóbbá) azáltal, hogy a körülményes nyelvspecifikus szakértői szabályok nélkülözhetetlenségének képzetét megcáfolják. A szakértői nyelvi szabályok megalkotása, korrigálása ugyanis magasan és speciálisan kvalifikált munkaerőt igényel, ami nemcsak a fejlesztési költségeket teszi magassá, de az alkalmazásfejlesztési időben is jelentős tételként jelenik meg, illetve számos hibalahetőséget rejt. Másrészt, a bemutatott adatvezérelt beszédfelismerési technikák más, morfológiájában és/vagy fonetikájában hasonló nyelvek esetén is jó eredményekkel kecsegtetnek. Ez részint a magyar kutatók számára adhat bíztatást a vizsgált nyelvek spektrumának szélesítésére, másfelől a magasabb nyelvi szintű technológiák (pl. gépi fordítás, információ-kivonatolás) számára szolgáltathat hazai alapot.

[18] felügyelet nélküli statisztikai módszerrel meghatározott morf lexikai egységekkel, n-gram statisztikai nyelvi modellel, triviális morf-graféma leképezéssel, graféma-szingleton akusztikus modellel

[19] súlyozott ejtésvariációk, kivételszótárak, graféma-fonéma átalakítási szabályok, fonetikai-fonológiai kategóriák.

Irodalomjegyzék

[Afify & Sarikaya+ 06] Afify, Mohamed; Sarikaya, Ruhi; Kuo, Hong-Kwang Jeff; Besacier, Laurent; Gao, Yuqing (2006): "On the use of morphological analysis for dialectal Arabic speech recognition", In INTERSPEECH-2006, pp. 1444-1447

[Aho & Hopcroft+ 74] Aho, Alfred V., John E. Hopcroft, and Jeffrey D. Ullman. 1974. The design and analysis of computer algorithms. Addison Wesley: Reading, MA.

[Allauzen & Mohri 02] Cyril Allauzen and Mehryar Mohri. On the Determinizability of Weighted Automata and Transducers. In Proceedings of the workshop Weighted Automata: Theory and Applications (WATA). Dresden, Germany, March 2002.

[Arisoy & Can+ 09] Ebru Arisoy, Dogan Can, Siddika Parlak, Hasim Sak and Murat Saraclar. Turkish Broadcast News Transcription and Retrieval. IEEE Transactions on Audio, Speech, and Language Processing, 17(5):874-883, July 2009

[Aubert 99] X. L. Aubert. One Pass Cross Word Decoding For Large Vocabularies Based On A Lexical Tree Search Organization. Proc. European Conf. on Speech Communication and Technology, pp. 1559–1562, Budapest, Hungary, September 1999.

[Bahl & Jelinek+ 83] L. R. Bahl, F. Jelinek, R. L. Mercer. A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 5, pp. 179–190, March 1983.

[Bahl & Brown+ 86] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, pp. 49–52, Tokyo, Japan, April 1986.

[Bahl & de Souza+ 93] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, M. A. Picheny. Context Dependent Vector Quantization for Continuous Speech Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 632–635, Minneapolis, USA, April 1993.

[Baker 75] J. K. Baker. Stochastic modeling for automatic speech understanding. In Reddy, R., editor, Speech recognition, pp. 512–542, New York, USA, Academic Press, 1975.

[Basu & Neti+ 99] S. Basu, C. Neti, N. Rajput, A. Senior. L. Subramaniam, A. Verma. Audio-visual large-vocabulary continous speech recognition in the broadcast news domain, IEEE Multimedia Signal Processing Conference (MMSP99), Denmark, Sept, 1999.

[Bauer 88] Bauer, W. 1988. On minimizing finite automata. EATCS Bulletin, 35.

[Baum & Eagon 67] L. E. Baum, J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of ecology. Amer. Math. Soc. Bull., Vol. 73, pp. 360–362, 1967.

[Bánhalmi & Kocsor+ 05] Banhalmi, A., Kocsor, A., Paczolay, D.: Magyar nyelvű diktáló rendszer támogatása újszerű nyelvi modellek segítségével, in Proc. of MSZNY 2005, pp. 337 – 347, Szeged, 2005.

[Bánhalmi & Paczolay+ 06] Banhalmi, A., Paczolay, D., Toth, L., Kocsor, A.: First Results of a Hungarian Medical Dictation Project, Proc. of IS-LTC 2006, pp. 23-26.

[Bellegarda & Nahamoo 90] J. R. Bellegarda, D. Nahamoo. Tied mixture continous parameter modeling for speech recognition. IEEE Trans ASSP, Vol. 38, No. 12, pp. 2033–2045, December 1990.

[Bellman 57] R. E. Bellman. Dynamic Programming. Princeton University Press, Princeton, USA, 1957.

[Berton & Fetter+ 96] A. Berton, P. Fetter, and P. Regel-Brietzmann. 1996. Compound words in large-vocabulary German speech recognition systems. In Proc. ICSLP, pp. 1165–1168, Philadelphia, PA, USA.

[Besling 94] S. Besling, "Heuristical and statistical methods for grapheme-tophoneme conversion," in KONVENS, Wien, Austria, Sep. 1994, pp. 23 – 31.

[Beulen & Ney 98] K. Beulen and H.Ney, Automatic Question Generation for Decision Tree Based State Tying, Proceedings of the ICASSP, pp- 805-808, Seattle, WA, 1998.

[Bisani & Ney 05] M. Bisani and H. Ney: Open Vocabulary Speech Recognition with Flat Hybrid Models. Proceedings of the European Conference on Speech Communication and Technology, Interspeech, pp. 725-728, Lisbon, Portugal, September 2005

[Bourlard & Morgan 93] Bourlard, H. and Morgan, N. (1993), "Continuous Speech Recognition by Connectionist Statistical Methods," IEEE Trans. on Neural Networks, vol. 4, no. 6, pp. 893-909.

[Brent 99] M. R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. Machine Learning, 34:71–105.

[Byrne & Hajic+ 01] W. Byrne, J. Hajic, P. Ircing, F. Jelinek, S. Khudanpur, P. Krbec, and J. Psutka. 2001. On large vocabulary continuous speech recognition of highly inflectional language—Czech. In Proc. Eurospeech, pp. 487–489, Aalborg, Denmark.

[Byrne & Doermann+ 04] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, W. J. Zhu, "Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 420-435, July 2004.

[Chen & Goodman 98] Stanley F. Chen and Joshua T. Goodman. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, 1998.

[Cohen 89] M. H. Cohen. Phonological structures for speech recognition. Ph.D. dissertation, University of California, Berkeley, USA, 1989.

[Creutz & Lagus 02] M. Creutz and K. Lagus. 2002. Unsupervised discovery of morphemes. In Proc. ACL/SIGPHON'02, pages 21–30.

[Creutz & Lagus 04] M. Creutz and K. Lagus. 2004. Induction of a simple morphology for highly-inflecting languages. In Proc. ACL/SIGPHON'04, pages 43–51.

[Creutz & Lagus 05a] Creutz, M. and Lagus, K., "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0.", Publications in Computer and Information Science, Report A81, Helsinki University of Technology, March, (2005)

[Creutz & Lagus 05b] Creutz, M. and Lagus, K., "Inducing the Morphological Lexicon of a Natural Language from Unannotated Text", In Proceedings of AKRR'05, Espoo, Finland, 15–17 June, (2005)

[Creutz & Hirsimäki+ 07] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, & A. Stolcke, Morph-based speech recognition and modeling of out-of-vocabulary words across languages, ACM Transactions on Speech and Language Processing 5(1), 2007.

[Czap 05] Czap L.: Audiovizuális beszédfelismerés és szintézis, PhD értekezés, BME, Budapet, 2005.

[Daniel 78] W. Daniel, Applied Nonparametric Statistics, Houghton Mifflin, 1978.

[Dempster & Laird+ 77] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal Royal Statistical Society, Series B, Vol. 39, No. 1, pp. 1–38, 1977.

[Erdoğan & Büyük+ 05] Erdoğan, H., Büyük, O., & Oflazer, K. (2005). Incorporating Language Constraints in Sub-word Based Speech Recognition. IEEE Automatic Speech Recognition and Understanding Workshop, 93–103.

[Geutner 95] Geutner, P. (1995). Using Morphology Towards Better Large-Vocabulary Speech Recognition System. IEEE International Conference on Acoustics, Speech and Signal Processing, Detroit, USA, 1995, 445–448.

[Glass 03] James R. Glass, A probabilistic framework for segment-based speech recognition, Computer Speech and Language 17 (2003) 137–152

[Goldsmith 01] Goldsmith, J. (2001). Unsupervised Learning of Morphology of Natural Language. Computational Linguistics, 27(2), 153–189.

[Good 53] Good, I.J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika, 40(3 and 4):237-264.

[Gordos & Takács 83] Gordos G., Takács Gy. (1983) *Digitális beszédfeldolgozás*, Műszaki Könyvkiadó, Budapest.

[Gósy 98] Gósy Mária. A zöngésségi hasonulás a (spontán) beszédben. Beszédkutatás 1998, Ed. Gósy Mária, Akadémiai kiadó, Budapest, pp. 1-20, 1998

[Gósy 04] Gósy Mária. Fonetika, a beszéd tudománya. Osiris kiadó, Budapest, 2004

[Haeb-Umbach & Ney 92] R. Haeb-Umbach, H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, pp. 13–16, San Francisco, USA, March 1992.

[Hain 02] T. Hain. Implicit pronunciation modeling in ASR. Proc. ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language, pp. 129–134, Estes Park, Colorado, USA, September 2002.

[Halácsy 06] Halácsy, P (2006). Benefits of deep NLP-based Lemmatization for Information Retrieval In: Working Notes for the CLEF 2006 Workshop, edited by Carol Peters . Cross Language Evaluation Forum.

[Hazen & Hetherington+ 02] Timothy J. Hazen, I. Lee Hetherington, Han Shu and Karen Livescu, "Pronunciation modeling using a finite-state transducer representation," Proceedings of ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation, Estes Park, Colorado, September, 2002

[Hermansky 90] H. Hermansky. (1990) Perceptual linear predictive (PLP) analysis of speech, *Journal of the Acoustical Society of America*, Vol. 87, No. 4, pp. 1738-1752.

[Hirsimäki & Kurimo 04] Teemu Hirsimäki and Mikko Kurimo: "Decoder Issues in Unlimited Finnish Speech Recognition", Proceedings of the 6th Nordic Signal Processing Symposium (Norsig 2004), June 9-11, 2004, Espoo, Finland, pp. 320-323.

[Hirsimäki & Creutz+ 06] Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J. (2006). Unlimited Vocabulary Speech Recognition with Morph Language Models Applied to Finnish. Computer, Speech & Language, 20(4), 515–541.

[Jelinek & Bahl+ 75] F. Jelinek, F. Bahl, R. L. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Trans. Information Theory, 21(3), pp. 250–256, 1975.

[Jelinek & Mercer 80] Jelinek, Frederick and Robert L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In Proceedings of the Workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands: North-Holland, May.

[Jurafsky & Ward+ 01] Jurafsky, Dan – Ward, Wayne – Jianping, Zhang – Herold, Keith – Xiuyang, Yu – Sen, Zhang. “What kind of pronunciation variation is hard for triphones to model?”, in IEEE ICASSP-01, Salt Lake City, Utah, 2001, pp. I.577–580.

[Kanji 94] G. Kanji, 100 Statistical Tests, SAGE Publications, 1994

[Kanthak & Ney 02] S. Kanthak, H. Ney. "Context-Dependent Acoustic Modeling Using Graphemes for Large Vocabulary Speech Recognition". In IEEE International Conference on Acoustics, Speech and Signal Processing, Vol 1, pp. 845-848, Orlando, FL, May 2002. download PostScript

[Kaplan & Kay 94] Kaplan, R. M. & Kay, M. (1994). 'Regular Models of Phonological Rule Systems'. Computational Linguistics 20, nr 3, 332-387.

[Katz 87] S. M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 35, No. 3, pp. 400–401, March 1987.

[Kirchoff & Vergyri+ 06] K. Kirchhoff, D. Vergyri, J. Bilmes, K. Duh, and A. Stolcke. 2006. Morphology-based language modeling for Arabic speech recognition. Computer Speech and Language, 20(4):589–608.

[Killer & Stüker+ 03] M. Killer, S. Stüker, and Tanja Schultz. Grapheme based Speech Recognition. Proc. Eurospeech, Geneva, Switzerland, September 2003

[Kurimo & Creutz+ 06] Kurimo, M., Creutz, M., Varjokallio, M., Arisoy, E. and Saraclar, Murat. “Unsupervised segmentation of words into morphemes - Morpho Challenge 2005, Application to Automatic Speech Recognition” In Interspeech 2006. Pittsburgh, USA, September 17-21, (2006).

[Kurimo & Puurula+ 06] Mikko Kurimo, Antti Puurula, Ebru Arisoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumae and Murat Saraclar. "Unlimited vocabulary speech recognition for agglutinative languages", In Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'06), 2006.

[Kwon & Park 03] O.-W. Kwon and J. Park. 2003. Korean large vocabulary continuous speech recognition with morpheme-based recognition units. Speech Communication, 39(3–4):287–300.

[Larson & Willett+ 00] M. Larson, D. Willett, J. Koehler, and G. Rigoll. 2000. Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. In Proc. ICSLP.

[Lee & Giachin+ 90] C.-H. Lee , E. Giachin , L. R. Rabiner , R. Pieraccini , A. E. Rosenberg, Improved acoustic modeling for continuous speech recognition, Proceedings of the workshop on Speech and Natural Language, p.319-326, June 24-27, 1990, Hidden Valley, Pennsylvania

[Leggetter & Woodland 95] C.J. Leggetter and P.C. Woodland. (1995) "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression, " Proc. ARPA Spoken Language Systems Technology Workshop, Austin, TX, pp. 110-115.

[Levinson & Rabiner+ 83] S. E. Levinson, L. R. Rabiner, M. M. Sondhi. An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell System Techn. Journal*, Vol. 62, No. 4, pp. 1035–1074, April 1983.

[Ljolje & Riley+ 99] A. Ljolje, M. Riley, and D. Hindle. The AT&T Large Vocabulary Conversation Speech Recognition System. In Proc. Eurospeech ’99, Budapest, Hungary, 1999.

[López & Graña+ 03] López, K., Graña, M., Ezeiza, N., Hernández, M., Zulueta, E., Ezeiza, A. and Tovar, C., “Selection of Lexical Units for Continuous Speech Recognition of Basque”, Proc. of CIARP, Havana, Cuba (2003) 244–250

[MacQueen 67] J. B. MacQueen. (1967) "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

[Mauuary 98] L. Mauuary. (1998) Blind Equalization in the Cepstral Domain for robust Telephone based Speech Recognition, *Proc. EUSPICO’98*, Vol.1, pp. 359-363.

[McDermott 97] E. McDermott, Discriminative Training for Speech Recognition, Ph.D. Thesis, Waseda Japan, 1997

[Mermelstein 76] P. Mermelstein. (1976) Distance measures for speech recognition, psychological and instrumental, *Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed.*, pp. 374–388. Academic, New York.

[Mohri & Sproat 96] Mehryar Mohri and Richard Sproat. An Efficient Compiler for Weighted Rewrite Rules.In 34th Meeting of the Association for Computational Linguistics (ACL '96), Proceedings of the Conference, Santa Cruz, California. Santa Cruz, California, 1996.

[Mohri 97] Mehryar Mohri. Finite-State Transducers in Language and Speech Processing. Computational Linguistics, 23:2, 1997.

[Mohri & Riley+ 98] Mehryar Mohri, Michael Riley, Don Hindle, Andrej Ljolje, and Fernando C. N. Pereira. Full Expansion of Context-Dependent Networks in Large Vocabulary Speech Recognition. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98). Seattle, Washington, 1998.

[Mohri & Pereira+ 02] Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. Weighted Finite-State Transducers in Speech Recognition. Computer Speech and Language, 16(1):69-88, 2002.

[Mohri & Riley 01] Mehryar Mohri and Michael Riley. A Weight Pushing Algorithm for Large Vocabulary Speech Recognition. In Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech '01). Aalborg, Denmark, September 2001.

[Myers & Rabiner, 81] C. S. Myers and L. R. Rabiner. (1981) A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7):1389-1409, September

[Nefian & Liang+ 02] A. Nefian, L. Liang, X. Pi, X. Liu, C. Mao, and K. Murphy. (2002) A coupled HMM for audio-visual speech recognition, *Proc. ICASSP, Orlando*, pp. 2013–2016.

[Németh & Mihajlik+ 07] Németh B., Mihajlik P., Tikk D., Trón V..: Statisztikai és szabály alapú morfológiai elemzők kombinációja beszédfelismerő alkalmazáshoz. MSZNY 2007: V. Magyar Számítógépes Nyelvészeti Konferencia, pp. 95-105, Szeged, 2007.

[Ney 84] H. Ney. The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 32, No. 2, pp. 263–271, April 1984.

[Ney & Mergel+ 87] H. Ney, D. Mergel, A. Noll, A. Paeseler. A Data-Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Vol. I, pp. 833–836, Dallas, USA, April 1987.

[Odell & Valtchev+ 94] J. J. Odell, V. Valtchev, P. C.Woodland, S. J. Young. A One-Pass Decoder Design for Large Vocabulary Recognition. Proc. ARPA Spoken Language Technology Workshop, pp. 405–410, Plainsboro, USA, March 1994.

[Olaszy & Németh+ 92] Olaszy G. - Németh G. - Gordos G.: The MULTIVOX multilingual text-to-speech converter. In: Bailly, G.-Benoit, C.-Swallis, T.(eds.): Talking machines: Theories, Models and Applications. Elsevier-North-Holland Publishers. Amsterdam 1992, 385-411.

[Olaszy & Németh+ 00] Olaszy G. - Németh G. - Olaszi P. - Kiss G. - Zainkó Cs. - Gordos G.: Profivox - a Hungarian TTS System for Telecommunications Applications. International Journal of Speech Technology. Vol 3-4. Kluwer Academic Publishers. 2000. 201-215

[Ordelman & Hessen+ 03] R. Ordelman, A. van Hessen, and F. de Jong. 2003. Compound decomposition in Dutch large vocabulary speech recognition. In Proc. Eurospeech, pp. 225–228, Geneva, Switzerland.

[Ortmanns & Ney+ 96] S. Ortmanns, H. Ney, A. Eiden. Language-Model Look-Ahead for Large Vocabulary Speech Recognition. Proc. Int. Conf. on Spoken Language Processing, pp. 2095-2098, Philadelphia, USA, October 1996.

[Povey & Woodland 02] D. Povey and P. Woodland, “Minimum phone error and I-smoothing for improved discriminative training,” in Proc. IEEE ICASSP, 2002,vol. 1, pp. 105–108.

[Prószéky & Tihanyi 93] Prószéky, Gábor; László Tihanyi: Humor: High-Speed Unification Morphology and Its Applications for Agglutinative Languages. La tribune des industries de la langue, No. 10. 28–29, OFIL, Paris, France (1993)

[Psutka & Ircing+ 05] Psutka, J., Ircing, P., Psutka, J. V., Hajic, J., Byrne, W. J., Mírovský, J.: "Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project", In INTERSPEECH–2005, (2005)1349–1352

[Psutka & Ircing+ 02] J. Psutka, P. Ircing, J.V. Psutka, V. Radova, V. Byrne, J. Hajic, S. Gustman and B. Ramabhadran, “Automatic Transcription of Czech Language Oral History in the MALACH Project: Resources and Initial Experiments,” in *TSD 2002*, Brno, Czech Republic, 2002, pp. 219-234.

[Psutka & Iljuchin + 03] J. Psutka, I. Iljuchin, P. Ircing, J. V. Psutka, V. Trejbal, W. Byrne, J. Hajic and S. Gustman, “Building LVCSR System for Transcription of Spontaneously Pronounced Russian Testimonies in the MALACH Project: Initial Steps and First Results,” in *TSD 2003*, Ceské Budejovice, Czech Republic, 2003, pp. 327-332

[Puurula & Kurimo 07] A. Puurula and M. Kurimo, “Vocabulary Decomposition for Estonian Open Vocabulary Speech Recognition,” in Proc. ACL-2007, pp. 89-95.

[Rabiner & Juang 93] Rabiner, L.R., and Juang, B-H. (1993) *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, New Jersey.

[Ramabhadran & Juang+ 03] Ramabhadran, B., Juang, J., and Picheny, M. “Towards Automatic Transcription of Large Spoken Archives - English ASR for the MALACH Project.” In International Conference on Acoustics, Speech, and Signal Processing, Genf (2003)

[Risannen 78] Rissanen, J. (1978), ‘Modeling By Shortest Data Description’, Automatica, Vol. 14, pp 465-471

[Schillo & Fink+ 00] C. Schillo, G. A. Fink, and F. Kummert, Grapheme based speech recognition for large vocabularies, in Int. Conf on Spoken Language Processing, Beijing, China, Oct. 2000, pp. 129 132.

[Schramm 06] H. Schramm. Modeling Spontaneous Speech Variability for Large Vocabulary Continuous Speech Recognition, Ph.D. dissertation, RWTH, Aachen, Germany, 2006.

[Schramm & Beyerlein 02] H. Schramm, P. Beyerlein. Discriminative Optimization Of The Lexical Model. Proc. ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language, pp. 105–110, Estes Park, Colorado, USA, September 2002.

[Siivola & Hirsimäki+ 03] Vesa Siivola, Teemu Hirsimäki, Mathias Creutz and Mikko Kurimo: "Unlimited Vocabulary Speech Recognition Based on Morphs Discovered in an Unsupervised Manner", Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH), pp. 2293-2296, 2003.

[Shafran & Hall 06] I. Shafran and K. Hall. 2006. Corrective models for speech recognition of inflected languages. In Proc. EMNLP, Sydney, Australia.

[Siivola & Pellom 05] Vesa Siivola and Bryan Pellom: "Growing an n-gram model", Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH), pp. 1309-1312, 2005

[Singh & Raj+ 99] Singh, R., Raj, B., Stern, R. M.: Automatic Clustering and Generation of Contextual Questions for Tied States in Hidden Markov Models. in Proc. Int. Conf. on Spoken Language Processing. Vol. 1 (1999) 117-120

[Siptár 95] Siptár Péter: Fonológia (Egyetemi jegyzet). Budapest: MTA Nyelvtudományi Intézet, 1995.

[Stolcke 98] A. Stolcke, “Entropy-based pruning of backoff language models,” in Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998, pp. 270–274.

[Stolcke 02] Stolcke, A., “SRILM – an extensible language modeling toolkit”, In Proc. Intl. Conf. on Spoken Language Processing, Denver (2002) 901–904

[Steinbiss & Tran+ 94] V. Steinbiss, B.-H. Tran, H. Ney. Improvements in Beam Search. Proc. Int. Conf. on Spoken Language Processing, Vol. IV, pp. 2143–2146, Yokohama, Japan, September 1994.

[Suontasuta & Hakkinen 00] J. Suontasuta and J. Hakkinen, “Decision tree based text-to-phoneme mapping for speech recognition,” in Int. Conf. on Spoken Language Processing, Beijing, China, Oct. 2000, pp. 199 – 202.

[Szarvas & Fegyó+ 00] M. Szarvas, T. Fegyo, P. Mihajlik, P. Tatai. Automatic Recognition of Hungarian: Theory and Practice. International Journal of Speech Technology, 3(3/4):237–251, 2000.

[Szarvas & Furui 02] Mate Szarvas and Sadaoki Furui "Finite-state transducer based Hungarian LVCSR with explicit modeling of phonological changes" Proc. ICSLP2002, Denver, U.S.A., pp.1297-1300 (2002-9)

[Szarvas 03] Máté Szarvas. "Efficient large vocabulary continuous speech recognition using weighted finite-state transducers - The development of a Hungarian dictation system" Ph.D. dissertation, TITECH, Tokyo, Japan, 2003.

[Szarvas & Furui 03a] Máté Szarvas and Sadaoki Furui. "Finite-state transducer based modeling of morphosyntax with applications to Hungarian LVCSR"
Proc. ICASSP2003, Hong Kong, China, vol.1, pp.368-371 (2003-4)

[Szarvas & Furui 03b] Szarvas, Mate / Furui, Sadaoki (2003): "Evaluation of the stochastic morphosyntactic language model on a one million word hungarian dictation task", In EUROSPEECH-2003, 2297-2300.

[Szarvas & Farkas 06] György Szarvas, Richárd Farkas, András Kocsor, A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms The Ninth International Conference on Discovery Science 2006,

[Szaszák & Vicsi 07] György Szaszák, Klára Vicsi: Using Prosody in Fixed Stress Languages for Improvement of Speech Recognition. COST 2102 Workshop 2007: 138-149

[Titterington & Smith+ 85] Titterington, D., A. Smith, and U. Makov (1985) "Statistical Analysis of Finite Mixture Distributions," John Wiley & Sons.

[Torkkola 93] K. Torkkola, “An efficient way to learn english grapheme-tophoneme rules automatically,” in IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Minneapolis, MA, April 1993, pp. 199 – 202.

[Tóth 06] Tóth, L.: Posterior-Based Speech Models and their Application to Hungarian Speech Recognition, Ph.D. Dissertation, University of Szeged, 2006.

[Tóth 09] Tóth, L.: Beszédfelismeresi kísérletek hangoskönyvekkel, Proc. MSZNY, pp. 206-216, 2009.

[Tóth & Kocsor+ 04] Tóth, L., Kocsor, A., Gosztolya, G.: Telephone Speech Recognition via the Combination of Knowledge Sources in a Segmental Speech Model, Acta Cybernetica, Vol. 16, No. 4, 2004.

[Trón & Németh+ 05] Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy. and Varga, D., “Hunmorph: open source word analysis”, In Proc. ACL 2005 Software Workshop, (2005) 77–85

[Trón & Halácsy+ 06] Viktor Trón, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda, and Eszter Simon (2006), Morphdb.hu: Hungarian lexical database and morphological grammar, In: Proceedings of 5th International Conference on Language Resources and Evaluation. ELRA, pages 1670--1673.

[Vicsi & Vig 98] Vicsi, K. - Vig, A.: Az elsõ magyarnyelvû beszédadatbázis, Beszédkutatás '98, MTA Nyelvtudományi Intézete, Budapest 1998, pp. 163-177

[Vicsi & Velkei+ 05] Vicsi K. Velkei Sz., Szaszák Gy., Borostyán G., Teleki Cs., Tóth Sz. L., Gordos G..: Középszótáras, folyamatos beszédfelismerőrendszer fejlesztési tapasztalatai, Proc. of MSZNY 2005, pp. 348 – 360.

[Vicsi & Tóth 02] Vicsi K., Tóth L. Kocsor A., Gordos G. Csirik J. (2002): MTBA - Magyar nyelvű telefonbeszéd adatbázis. Hiradástechnika 2002/8. sz. pp. 35-39.

[Vicsi & Szaszák 04] Klára Vicsi, György Szaszák: Examination of Pronunciation Variation from Hand-Labelled Corpora. TSD 2004: 473-480. Text, Speech and Dialogue, 7th International Conference, TSD 2004, Brno, Czech Republic, September 8-11, 2004, Proceedings. Lecture Notes in Computer Science 3206 Springer 2004, ISBN 3-540-23049-1.

[Vicsi et al.] Vicsi Klára et al. http://alpha.tmit.bme.hu/speech/databases.php

[Vintsjuk 68] T. K. Vintsyuk, „Speech discrimination by dynamic programming”, Kibernetika, Vol. 4, pp. 81-88, Jan.-Feb. 1968

[Wilcoxon 45] Wilcoxon, F. "Individual Comparisons by Ranking Methods." Biometrics 1, 80-83, 1945.

[Wittaker & Woodland 00] E. W. D. Whittaker and P. C. Woodland. 2000. Particle-based language modelling. In Proc. ICSLP, pp. 170–173, Beijing, China.

[Witten & Bell 91] Witten, Ian H. and Timothy C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory, 37(4):1085–1094, July.

[Young 06] S. J. Young. The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, Cambridge, England, December, 2006.

[Young & Odell+ 94] Young, S. – Odell, J. – Woodland, P. Tree-based state tying for high accuracy acoustic modelling. DARPA Human Language Technology Workshop, pages 307–312, March 1994.

[Zgank & Kacic+ 05] Zgank, A. - Kacic, Z. - Diehl F. - Juhar, J. - Lihan, S. - Vicsi, K. - Szaszák, Gy.: Graphemes as basic units for crosslingual speech recognition,, COST 278 Workshop, 2005

[Zsigri & Tóth+ 04] Zsigri, Gy., Toth, L., Kocsor, A. Sejtes, Gy.: Az automata és kézi szegmentálás ejtésvariációk okozta problémai, Proc. MSZNY 2004.

A tézispontokhoz kapcsolódó tudományos közlemények

Folyóiratcikkek

[J1] P. Mihajlik, Z. Tüske, B. Tarján, B. Németh, T. Fegyó: Improved Recognition of Spontaneous Hungarian Speech – Morphological and Acoustic Modeling Techniques for a Less Resourced Task, *IEEE Transactions on Audio Speech and Language Processing*, Volume 18, Issue 6, pp. 1588-1600, 2010.

[J2] P. Mihajlik, T. Révész, P. Tatai: Phonetic Transcription in Automatic Speech Recognition, *Acta Linguistica Hungarica*, Volume 49, Issues 3-4, pp. 407-425, 2003.

Cikkek szerkesztett könyvekben

[B1] P. Mihajlik, T. Fegyó, B. Németh, Z. Tüske, V. Trón: Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages: Hungarian ASR for the MALACH Project, In: V. Matousek, P. Mautner (ed.): *Text, Speech and Dialogue*, 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 2007, Proceedings, Lecture Notes in Computer Science, Volume 4629/2007, pp. 342-350.

[B2] Mihajlik P., Fegyó T., Tatai P.: Új eljárás a gépi beszédfelismerés környezetfüggő beszédhangmodelljeinek kialakítására. In: Gósy M (szerk.): *Beszédkutatás 2006*. MTA Nyelvtudományi Intézet, Budapest, 2006. pp. 218-230.

[B3] P. Mihajlik, P. Tatai, G. Gordos: Automatic Phonetic Transcription and Its Application in Speech Recogniser Training: A case study for Hungarian. In: P. Divenyi, S. Greenberg, G. Meyer (ed.): *Dynamics of Speech Production and Perception*, IOS Press, Amsterdam, NATO Science Series; I., 374., Life and Behavioural Sciences, 2006. pp. 245-262.

[B4] Mihajlik P., Tatai P.: Automatikus fonetikus átírás magyar nyelvű beszédfelismeréshez, In: Gósy M. (szerk.): *Beszédkutatás 2001*, MTA Nyelvtudományi Intézet, Budapest, 2001. pp. 172-185.

Konferenciacikkek[20]

[C1] B. Tarján and P. Mihajlik: On Morph-based LVCSR Improvements, *Proc. SLTU 2010*, May 3-5, 2010, Penang, Malaysia, pp. 10-16.

[C2] P. Mihajlik, B. Tarján, Z. Tüske, T. Fegyó: Investigation of Morph-based Speech Recognition Improvements across Speech Genres, *Proc. Interspeech 2009*, Sep. 6-10, 2009, Brighton, United Kingdom, pp. 2687-2690.

[C3] P. Mihajlik, T. Fegyó, Z. Tüske, P. Ircing: A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian, *Proc. Interspeech 2007*, August 27-31, 2007, Antwerp, Belgium, pp. 1497-1500.

[20] Minden külföldi cikk teljes terjedelmében lektorált. Az MSZNY cikkek absztrakt alapján lektoráltak.

[C4] Tüske Z., Mihajlik P., Fegyó T.: Spontán, nagyszótáras, folyamatos beszéd gépi felismerési pontosságának növelése beszélőadaptációval a MALACH projektben, *V. Magyar Számítógépes Nyelvészeti Konferencia*. 2007. december 6-7, Szeged, pp. 47-55.

[C5] Németh B., Mihajlik P., Tikk D., Trón V.: Statisztikai és szabály alapú morfológiai elemzők kombinációja beszédfelismerő alkalmazáshoz, *V. Magyar Számítógépes Nyelvészeti Konferencia*. 2007. december 6-7, Szeged, pp. 95-105.

[C6] Mihajlik P.: Koartikulációs modellek a magyar nyelvű gépi beszédfelismerésben, *IV. Magyar Számítógépes Nyelvészeti Konferencia*, 2006. december 7-8, Szeged, pp. 231-242.

[C7] T. Fegyó, P. Mihajlik, M. Szarvas, P. Tatai, G. Tatai: Voxenter™ – Intelligent Voice Enabled Call Center for Hungarian, *Proc. Interspeech 2003*, Sep. 1-4, 2003, Geneva, Switzerland, pp. 1905-1908.

[C8] T. Fegyó, P. Mihajlik, P. Tatai: Comparative Study on Hungarian Acoustic Model Sets and Training Methods, *Proc. Interspeech 2003*, Sep. 1-4, 2003, Geneva, Switzerland, pp. 829-832.

[C9] P. Mihajlik, T. Fegyó, P. Tatai, G. Gordos: Pronunciation Modeling in Continuous Number Recognition, *Proc. ECMCS 2001*, Sep. 11-13, 2001, Budapest, Hungary, pp. 330-333.

[C10] T. Fegyó, P. Mihajlik, P. Tatai, G. Gordos, Pronunciation Modeling in Hungarian Number Recognition, *Proc. Interspeech 2001*, Sep. 3-7, 2001, Aalborg, Denmark, pp. 1465-1468.

A szerző további tudományos közleményei (gépi beszédfeldolgozás témában)

Folyóiratcikkek

[J3] Németh G., Olaszy G., Bartalis M., Zainkó Cs., Fék M., Mihajlik P.: Beszédadatbázisok előkészítése kutatási és fejlesztési célok hatékonyabb támogatására. *Híradástechnika*, LXIII. évfolyam, 2008/5, pp. 18-24.

[J4] Tüske Z, Mihajlik P, Tobler Z, Fegyó T, Tatai P.: Beszéddetekciós módszerek vizsgálata és optimalizálása gépi beszédfelismerő rendszerekhez, *Híradástechnika*, LXI. évfolyam, 2006/3, pp. 59-67.

[J5] Szarvas M., Fegyó T., Mihajlik P., Tatai P.: Eredmények a magyar nyelvű nagyszótáras kapcsolt-szavas gépi beszédfelismerésben. *Híradástechnika*, LVI. évfolyam, 2001/6, pp. 31-36.

[J6] Szarvas M., Fegyó T., Mihajlik P., Tatai P.: Automatic Recognition of Hungarian: Theory and Practice, *International Journal of Speech Technology*, Volume 3, Numbers 3-4, pp. 237-251, 2000.

Cikkek szerkesztett könyvekben

[B4] Németh G., Olaszy G., Bartalis M., Kiss G., Zainkó Cs., Mihajlik P., Haraszti Cs.: Automated Drug Information System for Aged and Visually Impaired Persons, In: Miesenberger K, Klaus J, Zagler W, Karshmer A (ed.): *Computers Helping People with Special Needs*, Lecture Notes in Computer Science, Volume 5105/2008, Springer Berlin / Heidelberg, 2008. pp. 238-241.

[B5] Tüske Z., Simon M., Mihajlik P., Fegyó T.: Érzelmek automatikus felismerése a beszéd akusztikus jellemzői alapján. In: Gósy M. (szerk.): *Beszédkutatás 2007*. MTA Nyelvtudományi Intézet, Budapest, 2007. pp. 151-161.

[B6] Fegyó T., Mihajlik P., Tatai P.: Automatikus beszédfelismeréshez használt beszédhangmodellek betanítási módszereinek összehasonlító elemzése, In: Gósy M. (szerk.): *Beszédkutatás 2002*. MTA Nyelvtudományi Intézet, Budapest, 2002. pp. 185-196.

Konferenciacikkek

[C11] Tüske Z., Simon M., Mihajlik P., Gordos G.: A beszéd érzelemi töltetének számítógépes felismerése, *V. Magyar Számítógépes Nyelvészeti Konferencia*, 2007. december 6-7, Szeged, pp. 81-91.

[C12] Tarján B., Györki M., Mihajlik P., Gordos G.: Eredmények a magyar nyelvű beszédfelismerési konfidenciabecslésben. *IV. Magyar Számítógépes Nyelvészeti Konferencia*, 2006. december 7-8, Szeged, pp. 243-254.

[C13] Tüske Z., Mihajlik P., Tobler Z.: Új, zajbecsléssel kombinált beszéddetektálási eljárás a beszédfelismerési hatásfok javítására, *III. Magyar Számítógépes Nyelvészeti Konferencia*, 2005. december 8-9, Szeged, pp. 371-382.

[C14] P. Mihajlik, Z. Tobler, Z. Tüske, G. Gordos: Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech, *Proc. Interspeech 2005*, Sep. 4-8, Lisboa, Portugal, pp. 2677-2680.

[C15] Z. Tüske, P. Mihajlik, Z. Tobler, T. Fegyó: Robust Voice Activity Detection Based on the Entropy of Noise-Suppressed Spectrum, *Proc. Interspeech 2005*, Sep. 4-8, Lisboa, Portugal, pp. 245-248.